IJACSA
COMPUTER SCIENCE JOURNAL
International Journal of Advanced Computer Science and Applications

Volume 17 Issue 5
May 2026



ISSN 2156 - 5570 (Online)
ISSN 2158 - 107X (Print)

THE SCIENCE AND INFORMATION ORGANIZATION

www.ijacsa.thesai.org

Table of Contents

Paper 1: A Lifecycle-Oriented Taxonomy of Open-Source Tools for Machine Learning

Authors: Hamza Mallam Musa Mohammad, Isa Hussain Adam Mohammed, Abdullah Bajaber, Syed Abdur Rahman, Anas Sani, Atif Naseer

PAGES 1 – 12

Paper 2: Agentic Accountability: "The Buck Stops Where?" Ethical Frameworks for Human Oversight of Autonomous AI Systems

Authors: Ornella Bahidika

PAGES 13 – 22

Paper 3: ExtRA++: A Conceptual Architecture for a Deep Learning System for Aspect-Based Sentiment Analysis in User Reviews

Authors: G. Kanev, I. Valova

PAGES 23 – 29

Paper 4: Learning Analytics for Student Performance and Early Detection of At-Risk

Authors: Rhowel M. Dellosa

PAGES 30 – 38

Paper 5: An EWMA-Based Adaptive Thresholding Concept for Autoencoder-Based Concept Drift Detection in Data Streams

Authors: Siti Nurulain Mohd Rum, Qiao Song

PAGES 39 – 51

Paper 6: Hybrid CNN for Breast Cancer Detection in Multi-Modality Imaging

Authors: Emmanuel Ofotsu Kwesi Bannor, S. Sarah Maidin, Vinayakumar Ravi, Nguyen Thi Thu Thuy, Nghiem Thi-Lich

PAGES 52 – 60

Paper 7: A Strategic Cloud Adoption Readiness Index (SCARI) Framework: Empirical Development and Validation in the Ghanaian Context

Authors: Adade Sedom Percy, Emmanuel Ofotsu Kwesi Bannor, S. Sarah Maidin, Vinayakumar Ravi, Nguyen Thi Thu Thuy, Nghiem Thi-Lich

PAGES 61 – 68

Paper 8: Perceived Usefulness Without Ease of Use: A TAM-Based Evaluation of Enhanced Isabela State University Echague Online Resource App for a New Guild of Education (ISUE ORANGE) Functionalities

Authors: Ricardo Q. Camungao

PAGES 69 – 76

Paper 9: A Multilayer Secure Image Steganography Framework Using Edge-Adaptive Embedding and Pre-Encryption

Authors: A F M Zainul Abadin, Rossilawati Sulaiman

PAGES 77 – 87

Paper 10: Multimodal Machine Learning for Cybersecurity in Internet of Things Environments: A Literature Review

Authors: Abdelaaziz NASSIRI, Azeddine Khyat, Kama El Guemmat, Mohamed Aazi

PAGES 88 – 96

Paper 11: Optimizing the Multi-Omics Data Types and Variant Combinations for Accurate Breast Cancer Molecular Subtypes Classification

Authors: Sajid Shah, Azurah A Samah, Syed Hamid Hussain Madni, Sarina Sulaiman, Zuraini Ali Shah, Wong Yee Leng, Aryati Bakri

PAGES 97 – 107

Paper 12: An AI-Driven Framework for Software Effort Estimation Based on Developer Performance Metrics

Authors: Shaheer Ahmed, Nosheen Qamar, Faria Nazir, Nosheen Sabahat, Atif Ikram, Najla Abdulaziz Almousa, Hebah Abdullah Abubakr, Mohammed Abual-Rub, Abdulrahman Alojail, Marwan Abu-Zanona

PAGES 108 – 115

Paper 13: Cloud-Based Intelligent Surveillance for Digital Forensics: AI-Enhanced Criminal Investigations

Authors: Aseel Abdullah Aljuhani, Fatima Hamed Aljuhani

PAGES 116 – 123

Paper 14: A Fast and Efficient Residual Learning Framework Driven by Approximate Nearest Neighbor Search for Large-Scale Fingerprint-Based Visible Light Positioning

Authors: Huy Q. Tran, Huu Lam Phan, Tan Nguyen Van

PAGES 124 – 130

Paper 15: Insight Enhancement of Distortion Effects on the Quality in Imagery Environments

Authors: Hanan Hassan Ali Adlan, Mona Alanazi, Kolood Alenezi, Muneera Aldhabbah, Muneera Alsubai, Rafa Bahobail, Reem Alharbi

PAGES 131 – 145

Paper 16: A Hybrid Neuro-Fuzzy and Machine Learning Model for Student Performance Prediction in Online Mathematics Education

Authors: Beyza Esin Özseven, Turgut Özseven

PAGES 146 – 162

Paper 17: An Empirical Analysis of Google Play Data Safety Disclosures: A Consistency Study of Privacy Indicators in Mobile Gaming Apps

Authors: Bakheet Aljedaani

PAGES 163 – 175

Paper 18: Improving Emotion Recognition Accuracy Using a Multimodal Model (Face and Voice Video) Based on a Convolutional Neural Network (CNN)

Authors: Karnadi, Ermatita, Abdiansah

PAGES 176 – 185

Paper 19: Interlayer Coupling Strategy for Two-Layer Interdependent Command-and-Control Networks: Balancing Node Walk Betweenness and Operational Attributes

Authors: Bo Chen, Yangkun Wang, Yufeng Chen, Yuqi Yang

PAGES 186 – 197

Paper 20: Utilizing Structured Equation Modeling and Machine Learning in Investigating Digital Competency in Public School Teachers in Bukidnon, Philippines

Authors: Nathalie Joy G. Casildo

PAGES 198 – 207

Paper 21: DeepEdgeNet: An Edge-Cloud Deep Learning Framework for Efficient Environmental Monitoring in IoT Systems

Authors: Qamar H. Naith

PAGES 208 – 222

Paper 22: Machine Learning Evaluation of the HiTar-2024 Dataset for Intrusion Detection in Smart Manufacturing Environments

Authors: Adeeb Alhomoud

PAGES 223 – 232

Paper 23: ZK-FedMed: Privacy-Preserving Federated Learning for Cardiovascular and Renal Disease Prediction

Authors: Haewon Byeon

PAGES 233 – 244

Paper 24: An Interpretable Machine Learning Approach for Inflation Forecasting in Indonesia Using Domestic Macroeconomic Indicators

Authors: Ahmad Maimunif, Evaristus Didik Madyatmadja

PAGES 245 – 256

Paper 25: A Conceptual Model for Detecting Contactless Drug Distribution Based on Behavioural Analysis and Geospatial Visualisation

Authors: Medeu Kurmangali, Talgat Akimzhanov, Kanat Kazhibaev, Kulambayev Bakhytzhan, Moshkalov Altynbek

PAGES 257 – 269

Paper 26: Revisiting Support Vector Machines: A Distance-Based Alternative to Deep Learning for Efficient Text Classification

Authors: Siew Teng Koh, Anbuselvan Sangodiah, Norazira Binti A Jalil, Jafhate Edward, Nur Fatin Liyana Binti Mohd Rosely

PAGES 270 – 284

Paper 27: Pareto-Optimized Model Predictive Control for Dynamically Feasible Three-Dimensional Trajectory Generation in Robotic Manipulators

Authors: Zeinel Momynkulov, Sayat Ibrayev, Azizah Suliman, Yegenberdi Tenizbayev, Batyrkhan Omarov

PAGES 285 – 299

Paper 28: Enhanced IoT-Driven Load Forecasting with Metaheuristic-Optimized Deep Learning for Logistics Planning

Authors: Ramadan Babers, Walid Atwa, Mohamed Meselhy Eltoukhy, A. M. M. Madbouly

PAGES 300 – 309

Paper 29: A Smart IoT-Based Communication and Optimization System for Hajj and Umrah Services

Authors: T. H. F. Harumy, M. Gahara, M. F. Sipahutar, A. Zakiyah, N. A. Hasna, E. Ikhsan, Yusuf N. Buwono, Haposan Delon Pasaribu

PAGES 310 – 316

Paper 30: Blockchain Consensus Mechanisms Contributing to Improved Trust in Knowledge Sharing: A Systematic Review

Authors: Mohammad Fairus Bin Zulkifli, Rabiah Abdul Kadir, Mohamad Nazir Ahmad

PAGES 317 – 334

Paper 31: Generative AI in Higher Education: A Systematic Review with Emphasis on Programming and Computer Science Education

Authors: Abdulaziz Saidu Yalwa, Mohd Shahizan Othman, Lizawati Mi Yusuf, Muteb Sinhat Almarshadi

PAGES 335 – 347

Paper 32: Translating Job Advertisements into Competency Taxonomy: An Interpretable Approach for Robotics Recruitment Analysis

Authors: Zhiyan Xue, Yang Zhou

PAGES 348 – 363

Paper 33: Enhancing User Experience in Virtual Reality for Children with Autism: A Significant Review

Authors: Nur Aleesya Mohd Asri, Normala Rahim, Norsuhaily Abu Bakar, Wan Rizhan, Ismahafezi Ismail, Nur Saadah Mohd Shapri, Sarah Farhana Juhari

PAGES 364 – 376

Paper 34: Credit Card Fraud Anomaly Detection in Mobile Cloud Service Security Using Extended Isolation Forest with Hyperparameter Optimization

Authors: Nur Izura Udzir, Nur Farihin Bidin, Aliyu Usman Shehu, Madihah Mohd Saudi, Azuan Ahmad, Muhammad Harith Noor Azam, Shazrin Azlin Ruslan, Nor Azlinda Abdul Halim

PAGES 377 – 384

Paper 35: Fine-Grained Image Classification Using Vision Transformer Model

Authors: Zunaira Saleem, Uzma Jamil, Saman Iftikhar

PAGES 385 – 393

Paper 36: MRAE: Multi-Resolution Attention Ensemble with Hybrid CNN–Transformer Fusion for Breast Ultrasound Classification

Authors: Hemin Kareem Azeez Alshateri, Ahmed Harbaoui

PAGES 394 – 403

Paper 37: COAT: A Cross-Omics Attention Transformer for PAM50 Breast Cancer Subtype Classification

Authors: Soufiane El Atfa, Abdelmajid Hajami, Hamid Machhour, Hakim Allali

PAGES 404 – 415

Paper 38: Efficient Vulnerability Classification in IoT Networks: An Approach Using Convolutional Neural Networks and Tabu Search Optimization

Authors: Feras Fares AL-Mashagba, Mohammad Othman Nassar, Essam Said Hanandeh

PAGES 416 – 434

Paper 39: Scoping Review on Global Digital Policies and Transformation Strategies for Ageing Societies: Implications for Malaysia

Authors: Ahmad Wiraputra Selamat, Shamsul Arrieya Ariffin

PAGES 435 – 444

Paper 40: A Mid-Level Feature Fusion Framework Integrating PHLF and VGG16 for Robust Batik Pattern Detection

Authors: Jani Kusanti, Edi Noersasongko, Purwanto, Moch Arief Soeleman

PAGES 445 – 450

Paper 41: YOLO-CBAM: A Lightweight Attention-Guided Deep Learning Framework for Real-Time Road Damage Detection

Authors: Olzhas Olzhayev, Bakhytzhan Kulambayev, Azizah Suliman, Assel Rustem, Almira Madiyarova, Batyrkhan Omarov

PAGES 451 – 459

Paper 42: Deep Learning-Based Recommender System for Arabic Content with Integrated Sentiment Analysis of User Reviews

Authors: Amani Al-Ajlan, Nada Alshareef

PAGES 460 – 466

Paper 43: A Knowledge-Enhanced Cross-Modal Transformer Network for Sentiment Analysis in Intelligent Interaction

Authors: Chunyan Huang, Xinlu Sun

PAGES 467 – 478

Paper 44: Enhancing Traffic Congestion Forecasting with Explainable Deep Learning: A Framework Using LIME for Transparent Intelligent Transportation Systems

Authors: Ouhmidou Hajar, Nabou Abdellah, Elikram Moulay Ahmed

PAGES 479 – 486

Paper 45: Mammographic Image Classification Based on KAN-CBAM

Authors: Yuanyuan Wang, Vladimir Mariano

PAGES 487 – 498

Paper 46: Enhancing Virtual Team Performance Through a Quantitative-Based Management Model (VTPM)

Authors: Sathya A/L Nantha Kumar, Zulkefli Mansor

PAGES 499 – 509

Paper 47: An Explainable XGBoost-Based Framework for Robust Multi-Cohort Prediction of Pancreatic Cancer

Authors: Nada Ahmed El-Gammal, Rania Ahmed Abdel Azeem Abul Seoud, Sayed T. Muhammad

PAGES 510 – 519

Paper 48: A Hybrid PCA–Random Forest Model for Predicting Employee Performance from Work Ethics and Work Values

Authors: Nova E. Arquillano

PAGES 520 – 529

Paper 49: A New Approach for 3D Shape Retrieval Based on an Explainable Boosting Classifier

Authors: Fatima Rafii Zakani, Mohcine Bouksim, Khadija Arhid, Taoufiq Gadi, Mohamed Aboulfatah

PAGES 530 – 538

Paper 50: HBC-Glicko: A Bloom-Constrained Adaptive Assessment Architecture with Uncertainty-Aware Progression

Authors: Wissal EL Fougour, Mohamed Erradi

PAGES 539 – 552

Paper 51: A Secure Chaotic DNA-Based Framework for Satellite Image Encryption

Authors: Asim Seedahmed Ali Osman, Ibrahim Rizqallah Alzahrani, Aeshah Khalil I Alotaibi, Adil Mahmoud Mohamed Mahmoud, Daifallah Zaid Alotaibe

PAGES 553 – 562

Paper 52: A Diffusion-Based Generative AI Framework: An Exterior House Design from Textual Descriptions

Authors: Muhammad Amirul Akmal bin Ajusin, Noor Hasimah Ibrahim Teo, Rosniza Roslan, Raseeda Hamzah, Anita Ahmad Kasim

PAGES 563 – 570

Paper 53: Integration of Random Forest and Hough Transform for Cancer Classification Using Microarray Gene Expression Data

Authors: Hibah Alatawi, Hechmi Shili

PAGES 571 – 586

Paper 54: Detection of Unauthorized Use in SIEM Through Behavioral Analysis and Adaptive Rules

Authors: Julio Armando Landázuri Castro, Renato M. Toasa, Maryory Urdaneta Herrera

PAGES 587 – 594

Paper 55: Semantic Modeling of Medical Specialty Relationships Using Large Language Models

Authors: Ismail Bouajaja, Omar Elfahim, Omar Bouattane, Oussama Barakat, Abdelaziz Daaif

PAGES 595 – 602

Paper 56: An Intelligent Semantic-Aware Academic Library System Using Natural Language Processing and Knowledge Graphs

Authors: Mohamad Shady Alrahhal, Bandar Abdulwahed Fallatah, Omar F. Aloufi, Abdullah M. Barakeh

PAGES 603 – 612

Paper 57: Rice Pest Detection Using Enhanced YOLOv8n with Multilayer Contextual Attention and Deformable Snake Convolution

Authors: Shuangyuan Li, Jianglong Lin, Jiaming Liang, Tianyu Li

PAGES 613 – 623

Paper 58: Data-Driven Excimer LiDAR Framework for Joint Surface Reflectivity Mapping and Atmospheric Pollutant Profiling

Authors: Sandugash Dospanbetova, Gulzat Ziyatbekova, Murat Baktybayev, Botakoz Smagul, Yermakhan Zhabayev, Zhanar Bidakhmet

PAGES 624 – 635

Paper 59: Cervical Cytology Classification Using Multiple CNN Architectures with Transformer-Based Feature Enhancement

Authors: Mehreen Sirshar, Omama Shakeel, Nayyab Asim, Fakeeha Jafari, Hani Almoamari, Adnan Nadeem, Mohammad Zubair Khan, Ibrahim Aljubayri

PAGES 636 – 646

Paper 60: Methodology for Identifying Soiling on PV Panels Using RGB Images and Deep Learning

Authors: Katerina Gabrovska-Evstatieva, Tsvetelina Kaneva, Irena Valova, Dimitar Trifonov, Nikolay Valov, Ventsislav Keseev, Nicolay Mihailov, Boris Evstatiev

PAGES 647 – 655

Paper 61: Gas-Efficient Smart Contract Design: Quantifying Refactoring Impact on EVM Execution Costs

Authors: Nur Haliza Abdul Wahab, Juniardi Nur Fadila, Nur Faszha Razali, Keng Yinn Wong

PAGES 656 – 668

Paper 62: Agent-Oriented Fuzzy Decision Support System for Multi-Criteria Evaluation of Sustainable Investments in the Agro-Industrial Sector

Authors: Aigerim Omurtayeva, Ulzhan Makhazhanova, Akmaral Kulamanova, Dinara Kargabaeva, Bolat Tassuov, Adilbek Tanirbergenov

PAGES 669 – 679

Paper 63: Phonetic Completeness Over Prosodic Diversity: Syllable-Level Synthetic Corpus Construction for Low-Resource Penang Hokkien Speech Synthesis

Authors: Yu Liang Lai, Yen Min Jasmina Khaw, Seng Poh Lim, Tien Ping Tan

PAGES 680 – 690

Paper 64: Exploring New Possibilities: Virtual Reality in Supporting Children with Autism

Authors: Brayan J. Lozano-Martel, Eduardo F. Araujo-Vasquez, Segundo E. Cieza-Mostacero

PAGES 691 – 698

Paper 65: Augmented Reality in Automotive Technical Training: Technological Innovation Through Mobile Applications in Trujillo

Authors: Julissa B. Polo-Holguin, Eduardo F. Araujo-Vasquez, Segundo E. Cieza-Mostacero

PAGES 699 – 708

Paper 66: CFD-Based Aerothermal Design Verification of Coffee Cabinet Dryer

Authors: Gerry M. Castillo, Aeron R. Mojica, Jhon Heron P. Carsocho, Ray Rojan C. Romeroso, Sheryl Dinglasan-Fenol, Al Eugene L. Torres, Gee Jay C. Bartolome

PAGES 709 – 718

Paper 67: Unveiling Temporal Dynamics of Consumer Emotions in Online Reviews: An Emotion Mining Approach Using Ekman's Emotion Model

Authors: Azra Shamim, Abdulwahab Ali Almazroi

PAGES 719 – 730

Paper 68: Privacy Preserving Federated Graph Learning with Data Envelopment Analysis Driven Interpretable Customer Segmentation Framework

Authors: Pentareddy Ashalatha, G. Krishna Mohan

PAGES 731 – 743

Paper 69: Explainable Cognitive Graph Intelligence Framework for Multi-Task Graduate Employability and Salary Prediction

Authors: Hameeda Khatoon, Chandra Prakash Vudatha

PAGES 744 – 758

Paper 70: Large Language Models for Arabic Automated Essay Scoring

Authors: Leena Najjar, Liyakathunisa Syed

PAGES 759 – 769

Paper 71: Revolutionizing Urban Waste Systems: A Comprehensive Review of IoT, AI, and Optimization Techniques for Smart Waste Management

Authors: Lahcen GOUSKIR, Abdelmoula ABOUHILAL, Mohamed GOUSKIR, Mohamed BASLAM, Soufiane BELHOUIDEG, Hanaa Hachimi

PAGES 770 – 777

Paper 72: Deep Learning-Enhanced Excimer Laser LiDAR System for High-Resolution Earth Surface Monitoring and Multi-Platform Validation

Authors: Sandugash Dospanbetova, Gulzat Ziyatbekova, Murat Baktybayev, Botakoz Smagul, Yermakhan Zhabayev, Zhanar Bidakhmet

PAGES 778 – 788

Paper 73: Federated and Cross-Domain Student Performance Prediction

Authors: Sam Zhe Xuan, P. Ganesh Kumar, C. Rani, Kanagalakshmi, R. RajiniGanth, Atif Mahmood

PAGES 789 – 798

Paper 74: Intelligent Traffic Surveillance: Machine Learning-Based Detection of Vehicle Speed Violations

Authors: Niloy Kanti Paul, Dipanwita Saha, Kaushik Biswas, Sultanul Arifeen Hamim, Tanvir Ahmed, Rifath Mahmud

PAGES 799 – 810

Paper 75: Malak: A Python Toolkit for Edge AI Model Optimization

Authors: Mohammed Hassan Alnemari

PAGES 811 – 818

Paper 76: RGB-D Bin-Picking System for Ergonomic Automotive Clip Assembly: 3D Annotation, Deep Learning, and 6-DOF Pose

Authors: Brahim Bergor Beguiel, Ibrahim Hadj Baraka, Yassir Zardoua

PAGES 819 – 833

Paper 77: Urdu-ClimateGPT: Adapting LLM for Climate Data in Urdu Language

Authors: Muhammad Farooq, Muhammad Asif Habib, Jabeen Sultana, Muhammad Umar Aftab

PAGES 834 – 843

Paper 78: Leveraging MCESTA and Large Language Models for Next-Generation Similarity Learning

Authors: Mohamedou Cheikh Tourad, Abdelmounaim Abdali, Mohamed Dhleima, Naoual Mouhni, Sana Chakri, Ibtissam Amalou, Saadbouh Cheikh El Mehdy

PAGES 844 – 853

Paper 79: Enhanced Pied Kingfisher Optimization Algorithm with Hovering Scouts and Foraging Flocks Mechanisms

Authors: Aarhus Dela Cruz

PAGES 854 – 869

Paper 80: A Lightweight Explainable Hybrid Deep Learning Approach for Early Skin Disease Detection

Authors: Mohammad Barr

PAGES 870 – 879

Paper 81: Intrusion Detection System for Ransomware in Network Traffic Using Supervised Machine Learning

Authors: Ziad Almulla, Moath Alamri, Mounir Frikha

PAGES 880 – 897

Paper 82: Domain Knowledge-Enhanced Welding Robot Path Planning Algorithm

Authors: Ming Liu, Peng Shao

PAGES 898 – 908

Paper 83: Optimizing Whiteboard Digitization with Clarix: An Automated Vision-Based System for Real-Time Text Extraction and Surface Cleaning

Authors: Abdellah NABOU, Ameksa Mohammed, Ezzahoud Hajar, Bazgour Yassine, Abouhane Zahra

PAGES 909 – 916

Paper 84: Adaptive Temporal Windowing for Streaming Outlier Detection Under Dynamic Arrival Rates

Authors: Hend Maher, Mohamed Khafagy, Heba Nagaty

PAGES 917 – 925

Paper 85: Deep Learning-Based Road Damage Detection Using Improved YOLOv8: Model Performance and Implementation

Authors: Aulia Rahman, Hwa Jen Yap, Rusdha Muharar

PAGES 926 – 934

Paper 86: Smarter Bridges: Leveraging Artificial Intelligence to Reshape University-Industry Technology Transfer

Authors: Khaouja Mohammed, Sanaa DFOUF, Kaoutar ERRAKHA, Hanan ELHARISSI, Fekkak HAMDI

PAGES 935 – 949

Paper 87: Development of an Arabic Pet Adoption System with Hybrid Recommendations

Authors: Hailah Alballaa

PAGES 950 – 959

Paper 88: State-AttentNet: A Dynamic Volatility-Adaptive Hybrid Framework for Market State Classification in Frontier Economies

Authors: Md. Abul Kalam Azad, Abdul Kadar Muhammad Masum, Najmus Saadat, Esrat Jahan, Ramona Birău, Virgil Popescu, Iuliana Carmen Bărbăcioru, Ștefan Mărgăritescu

PAGES 960 – 970

Paper 89: Cognitively Aligned Assessment Item Generation with Open-Source LLMs: A Comprehensive Evaluation on LearningQ

Authors: Mahmoud Badry, Walaa Medhat, Shereen A. Taie, Asmaa Hashem Sweidan

PAGES 971 – 979

Paper 90: A Machine Learning Approach for Targeting Lucrative Customer Segments

Authors: Leen Almajed, Haifa Almakhdhoub, Jana Alzeydan, Alaa Bin Sleem, Ouiem Bchir

PAGES 980 – 990

Paper 91: A Human-Centered Evaluation of AI-Generated Guidance: Integrated Statistical and Machine Learning Analysis with a Risk Framework for High-Stakes Domains

Authors: Omar Al-Turki, Felwah Alqahtani, Eman Alqahtani, Sarah Alswedani, Sami Alshmrany, Rashid Mehmood

PAGES 991 – 1006

Paper 92: D-HAN-Net: A Hybrid Dual-Stream Architecture for Corporate Bankruptcy Prediction via Multimodal Fusion

Authors: Md. Abul Kalam Azad, Abdul Kadar Muhammad Masum, Najmus Saadat, Esrat Jahan, Ramona Birău, Virgil Popescu, Iuliana Carmen Bărbăcioru, Ștefan Mărgăritescu

PAGES 1007 – 1019

Paper 93: A Reliability-Aware Visual-Inertial Odometry for Dynamic and Low-Texture Environments

Authors: Yelu Liu, Ruokun Qu, Mengcheng Xu, Chenglong Li, Hui Jiang

PAGES 1020 – 1030

Paper 94: Trustworthy App Detection in Saudi Mobile Finance: Bridging Deep Learning and Interpretability

Authors: Raed Alharbi, Maryam Alghamdi

PAGES 1031 – 1041

Paper 95: Benchmarking Deep Learning for PM2.5 Forecasting in IoT-Based Smart Cities: GCN Spatial Encoding and Transformer Temporal Modeling

Authors: Abdessamad BADOUCH, Kaoutar BELHOUCINE

PAGES 1042 – 1049

Paper 96: A Controlled Benchmark of Open-Source and Proprietary LLMs for Few-Shot Microbiome IBD Classification

Authors: Nouhaila En Najih, Soufiane Hamida, Ahmed Moussa

PAGES 1050 – 1058

Paper 97: A Lightweight and Robust APBT–LBP Deep Feature Zero-Watermarking Framework with DNA Encryption for Medical Images

Authors: Ranjan Kumar Senapati, Prasanth Mankar, B Padmaja, Chilamakuru Nagesh, Pradeep Kumar, Gandikota Ramu, Gandharba Swain

PAGES 1059 – 1074

Paper 98: A Privacy-Aware Federated Hybrid Model for Multimodal Mental Health Analysis

Authors: Yusra, Riaz UlAmin

PAGES 1075 – 1084

Paper 99: Design and Evaluation of a Ubiquitous Learning Game for Adolescent E-Cigarette Addiction Prevention

Authors: Lamyae Bennis, Khalid Kandali, Hamid Bennis

PAGES 1085 – 1094

Paper 100: MILP for Multimodal Urban Transport: Formal and Informal Sectors

Authors: Oumar KONE, Yapi Fiacre Aristide EDI, Kouassi Hilaire EDI, Pawoumodom Matthias TAKOUDA

PAGES 1095 – 1107

Paper 101: Artificial Intelligence in Motorsport: A Scoping Review of Applications, Challenges, and Research Gaps

Authors: Radu C. Lucaciu, Cornelia Aurora Győrödi

PAGES 1108 – 1119

Paper 102: Benchmarking Deep Vision Architectures: From Controlled Datasets to Real-World Clinical Validation

Authors: Raghda Essam Ali, Noha Ahmed Saad El-Dien, Magi Hossam Eldin Mahfouz

PAGES 1120 – 1129

Paper 103: Smart IoT Framework for Sustainable Cities: Integrating Information Systems and Artificial Intelligence

Authors: Amarildo Rista, Alma Stana

PAGES 1130 – 1145

A Lifecycle-Oriented Taxonomy of Open-Source Tools for Machine Learning

Hamza Mallam Musa Mohammad[1], Isa Hussain Adam Mohammed[2], Abdullah Bajaber[3],
Syed Abdur Rahman[4], Anas Sani[5], Atif Naseer[6]
University of London, London[1]
Data Science Program, University of Colorado Boulder, Boulder, CO, USA[2]
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, 21421, Saudi Arabia[3]
Jawaharlal Nehru Technological University, Hyderabad, India[4]
Goldsmiths, University of London[5]
Deanship of Postgraduate Studies and Research, Umm Al-Qura University, Makkah, 21955, Saudi Arabia[6]

***Abstract*—Machine learning (ML) offers various tools, frameworks and platforms for resolving complex problems in computational science and engineering. Machine learning frameworks have emerged as the cornerstone of modern research and innovation. It redefines how knowledge is produced, validated, and disseminated. Open-source machine learning frameworks are emerging as a promising way to solve the challenges of large datasets, real-time constraints and heterogeneous system components. This study provides an extensive overview of open source tools based on the ML lifecycle. These tools are evaluated based on their purpose and key features for each stage of lifecycle, assisting researchers and practitioners in making informed decisions according to their requirements. The key challenges are identified and future research directions are also outlined.**

***Keywords*—Open-source; lifecycle; taxonomy; machine learning; challenges**

I. Introduction

Recent developments in AI and ML, especially in their capabilities to automate complex operations, uncover hidden patterns from large amounts of data and help in predictive analysis. This insight is valuable in making decisions and have a significant impact on several sectors like the healthcare, finance, manufacturing and entertainment [1], [2], [3]. The core factors for these advances include open source solutions that ease the creation, development, testing and deployment of AI applications and ML models [4], [5], [6], [7], [8].

Open-source tools help to improve development, as well as make advanced analytics more accessible for users with different experience levels [9], [11], [10]. Nowadays, there is a variety of AI/ML tools available in the market to fulfill diverse needs and preferences of their potential customers [13], [12]. The available resources include fundamental ML libraries such as TensorFlow, PyTorch and high-level frameworks like Keras or Scikit-learn. In addition, complementary components enable structured knowledge representation, rule-based reasoning, uncertainty modeling, sequential decision-making, and integration with real-world systems such as robotics platforms. Furthermore, these tools offer essential functionalities that cover such areas as data pre-processing, features selection, algorithm choice, optimization, inference, model evaluation, deployment and help in achieving high performance in AI models [14], [15]. Despite the variety of available options, choosing an adequate tool is still difficult. It should be noted that the choice of the right tool depends on the type of problems, complexity of data, licensing issues and skills of the practitioner [7], [16], [17], [18].

The adoption of open-source artificial intelligence (AI) and machine learning (ML) frameworks has become fundamental to current computational research and innovation endeavors. In addition to economic benefits that open source technology offers, openness, collaboration and availability are among its most important features. These qualities align well with sound scientific research. Since licensing restrictions are not applicable to open source software frameworks, researchers, entrepreneurs, and other public institutions can use them for collecting data, processing, computations and model development. The open access to source code, model designs and training steps also improves reproducibility, which is the foundation of scientific research integrity.

Collective intelligence of global open-source communities works as an ongoing peer-review system. It is helpful for improvement of tools via prompt updates, bug fixes and new features. It is worth noting that such processes usually occur much faster compared to similar operations in closed proprietary projects. These community efforts promote best practices via providing quality documents, benchmarks and educational resources. The common formats and APIs help integrate various solutions within different organizations. Version history help in detecting early stages problems and facilitates track changes. Researchers and developers can reuse code, pipelines and shared libraries to match specific needs.Public availability of code enables a broad community of experts to identify bugs, detect vulnerabilities, and improve overall software reliability. Open-source tools are also useful in resource-constrained environments for developing local AI expertise. This approach facilitates global participation and provides more diverse ideas. Open-source solutions are very beneficial in improving governance and cyber-security. This aspect is particularly relevant in healthcare, transport, and climate science for continuous research growth in AI and machine learning.

This study presents a detailed review of the open-source ML tools by considering entire lifecycle based on SLR methodology [19], [20]. We focus on the features, advantages,

and limitations of these tools. This study aims to provide useful insights regarding the analysis of such tools and helping individuals in decision-making based on individual needs. The contributions of this research work are as follows:

- A detailed analysis of different stages of the machine learning lifecycle is provided.
- A lifecycle-based taxonomy is introduced for open source machine learning tools, which provide insight to different open source ML tools based on their functionalities. We explore their key features, use cases, and other distinguishing characteristics.
- A detailed discussion and comparative analysis is established on widely used ML tools, emphasizing their unique features and suitability for different types of machine learning applications.
- The key challenges, risks, and future directions are identified, providing insights into potential areas for improvement and innovation.

The remainder of this study is organized as follows: Section II present research methodology and initial analysis. Section III presents an end-to-end view of the machine learning lifecycle establishing the foundation for our analysis. Section IV introduces a lifecycle-oriented taxonomy to organize open-source tools by functionality, summarizes their capabilities, and use cases. Section V provides a comparative analysis and discusses trade-offs and selection guidance for different application contexts. Section VI outlines key challenges, risks, and future research direction. Section VII concludes with a synthesis of key findings and practical implications.

II. Research Methodology and Initial Analysis

This study systematically reviews recent publications in the field of open-source machine learning (ML) tools. We investigate open-source ML tools across different stages of the ML lifecycle, including problem framing, data collection and pre-processing, feature engineering, model development and training, deployment, monitoring, and governance. The investigation draws on sources from prominent academic research platforms and technical repositories, including ACM Digital Library, Elsevier, Google Scholar, IEEE Xplore, SpringerLink, and GitHub. All relevant publications and documentation available up to June 2025 were reviewed, with a primary focus on studies published from 2021 onward.

To ensure comprehensive coverage of the domain, two categories of keywords were used. The first set, including "ML lifecycle", "machine learning lifecycle", "ML lifecycle taxonomy", and "ML system engineering" was used to establish the scope of the overall investigation. The second set, comprising "open source ML tools", "data lifecycle open source tools", "deep learning open source tools", "development open source tools", "model deployment open source tools", "MLOps and workflow open source tools", "model serving and integration open source tools", "monitoring and governance open source tools" was used to refine the search toward specific lifecycle stages and functional tool classes. In the context of production deployment and trustworthiness, additional qualifiers were incorporated, including "privacy in machine learning", "supply-chain security", "fairness assessment", and "regulatory compliance".

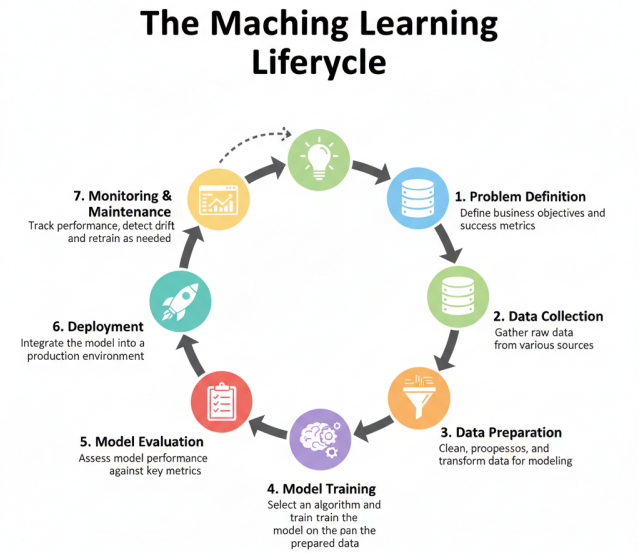


Fig. 1. Machine learning lifecycle.

At the beginning of the review process, an initial pool of research articles was identified based on relevance to the broad domain of open-source ML tools and frameworks. Following a rigorous screening process involving duplicate removal, abstract filtering, and full-text assessment, few publications were identified as directly aligned with the scope of this study.

III. Machine Learning Lifecycle

To provide a comprehensive framework for analyzing open-source tools, this section will explain how the machine learning lifecycle works in various stages. The machine learning lifecycle involves several stages which are part of the systematic process in the development of the models [21]. A brief analysis of the machine learning lifecycle along with the evaluation of current applications of machine learning algorithms are described [22]. Fig. 1 presents the machine learning lifecycle.

A. Problem Framing

At the core level, the machine learning lifecycle starts with identifying a problem. In this phase, a problem is identified and its structure is defined [23]. It formulates objectives and performance metrics that ensure consistency between model goal and the company's objectives based on availability of data and resources. This is a joint effort between business experts and data scientists to determine the expected outcomes [24]. Ethical factors such as fairness, harms, and threats are very important and need to be taken into consideration at early stages because they greatly affect data collection, model creation and implementation subsequently [21]. The criteria for success should be more than accuracy and extend to robustness and reliability as well.

B. Data Collection

This stage focuses on acquiring raw data that are used for model training [24]. It involves collecting data samples

through observation and measurement of the real-world system, process, or phenomenon for which an ML model needs to be built [25], [26]. It is critical in the lifecycle of machine learning models, as the quality and relevance of the gathered information directly influence the model performance.

Data can be collected from a range of internal sources, such as organizational databases, logs, and historical records, as well as external sources such as open-source platforms like Kaggle or the UCI Machine Learning Repository, social media streams for sentiment analysis, and market research studies. It is important to ensure that the selected data are relevant to the defined problem. Data may take various forms, including structured formats stored in relational databases, unstructured content such as text, images, or audio, and semi-structured formats such as JSON or XML. It is also important to determine the appropriate amount of data, balancing volume and quality to avoid issues like under- or overfitting while considering computational constraints [21]. The quality and relevance of collected data is equally critical, as these factors directly impact model performance and validity of the predictions.

C. Data Preparation

Data preparation is an important stage in the machine learning lifecycle. At this stage, the raw data is transformed into a clean, organized and suitable format for model training and evaluation [27], [28], [29]. The data cleaning involves handling missing values through deletion or imputation, identifying and managing outliers using statistical techniques and removing duplicates to maintain data integrity. Once the dataset is clean, data transformation techniques are applied. The data transformation techniques may include normalization, standardization to scale numerical features appropriately and encoding methods (one-hot or label encoding) to convert categorical data into machine-readable formats. In data splitting, dataset is partitioned into training, validation and testing sets. It helps in model evaluation and prevents overfitting. The advanced data engineering practices, such as pre-processing have streamlined the exploratory data analysis (EDA) and data preparation stages [30].

D. Feature Engineering

Feature engineering plays a significant role in improving model performance by creating new meaningful features. It may involve mathematical operations and the selection of only the most relevant features using methods such as recursive feature elimination, statistical tests, or feature importance measures [31], [32]. Proper data splitting is also essential, dividing the dataset into training, validation and test sets to support model training, hyperparameter tuning, and unbiased performance evaluation. In cases involving image, text, or other unstructured data, data augmentation techniques can be applied to expand the training data set artificially through modifications such as rotations, noise addition, or text perturbations or to further add data samples to the dataset [34], [33].

Data normalization and feature scaling techniques may also be applied to ensure that data adhere to a uniform scale and facilitate the learning process [24].

E. Model Selection, Training and Optimization

The selection of an appropriate ML model depends on various factors, including the nature of the problem (classification, regression, etc.), the characteristics of the data (structured, unstructured), and the available computational resources. Popular choices include decision trees, support vector machines, neural networks, and ensemble methods [24].

Model training plays a pivotal role in the machine learning lifecycle. In this phase, the selected algorithm learns patterns from the prepared data set by adjusting its parameters to reduce prediction errors. Performance metrics such as loss and accuracy are monitored throughout to ensure proper convergence and detect issues such as overfitting. Hyper-parameters are high-level configuration settings of the ML algorithm that can significantly impact its learning capacity. In addition to this, key hyper-parameters such as the learning rate, the number of training epochs, and the batch size are configured to control the model's learning process. Optimization algorithms such as grid search, random search, bayesian optimization, stochastic gradient descent, or Adam iteratively update the model parameters to minimize loss [24].

F. Model Evaluation and Validation

After training, the model is validated on a separate validation set to evaluate its generalization capability and tune hyper-parameters if required [35], [36]. It ensures that the model generalizes well beyond the training dataset and meets the intended objectives. The process begins with the selection of appropriate evaluation metrics based on the type of problem [24]. Once metrics are defined, the model is assessed using a validation set.

Model evaluation involves checking for overfitting or underfitting by comparing training and validation performance. Overfitting occurs when the model performs well in training data but poorly in validation data, while underfitting indicates that the model is too simple to capture underlying patterns. Sensitivity analysis can be conducted to study how variations in input features influence predictions, revealing the model robustness and identifying critical features. By systematically evaluating the model, practitioners ensure that its performance is reliable, generalizes effectively and is suitable for real-world deployment.

G. Model Deployment

The deployment of the model is the decisive stage of the machine learning lifecycle, where the trained model is integrated into a production environment to generate predictions on new real-world data [24]. This phase ensures that the model operates efficiently, reliably and securely. The process begins with selecting an appropriate deployment method, such as batch processing for scheduled or real-time inference for applications that require immediate responses through APIs. The deployment environment may involve cloud platforms like AWS, Azure, or Google Cloud, chosen for their scalability, services and global reach. The model is then serialized into a suitable format, such as Pickle or TensorFlow, so that it can be easily loaded and executed in production pipeline. If the model needs to be accessed by external applications, an API is developed using frameworks like FastAPI or Flask. Once deployed,

continuous monitoring becomes critical to track performance metrics such as prediction accuracy, latency and resource usage. A structured plan for model updates helps maintain long-term performance by incorporating retraining schedules, version control and user feedback. The post-deployment performance evaluation ensures that the model continues to meet business goals.

H. Monitoring and Maintenance

Monitoring and maintenance are critical stages of the machine learning lifecycle, ensuring that the models remain reliable, accurate and aligned with real-world requirements. Once a model is operational, continuous performance monitoring becomes essential to detect degradation timely. This includes tracking key metrics (accuracy and precision) and system-level indicators (latency and throughput). The detection of data drifts is equally important, as it can significantly impact predictions. Logging and auditing play an important role in maintaining transparency and accountability by recording inputs, outputs, errors and model updates. User feedback also contributes to continuous improvement, offering practical insights into the behavior of the real-world model. In addition, regular reviews help address technical debt by refactoring code, updating dependencies and improving infrastructure reliability. The documentation must be kept up-to-date to reflect model updates, deployment changes and operational guidelines. The techniques such as model explainability, bias mitigation and transparency provide insight into model behavior, allowing stakeholders to assess model performance and compliance with regulatory and ethical standards [30]. Finally, a well-defined incident response plan enables teams to quickly mitigate issues related to performance or security breaches. Together, these monitoring and maintenance practices create a proactive framework that keeps machine learning models reliable and effective throughout their lifecycle.

IV. Lifecycle-Oriented Taxonomy of Open-Source Machine Learning Tools

The existence of a variety of open source machine learning tools enables a reliable, scalable, and reproducible data workflow throughout the machine learning pipeline. A structured taxonomy for machine learning experiment tracking is introduced and applied to widely used tools [37]. A comprehensive taxonomy of open source machine learning tools is presented in Fig. 2. It includes open source machine learning tools for data engineering, deep learning, model deployment and optimization, workflow orchestration, model serving and system integration, monitoring and governance, development, visualization and knowledge management.

A. Data Lifecycle

Data operations in machine learning pipelines include ingestion, distributed computing, quality and validation, lineage and governance, labeling and annotation, feature engineering, data augmentation, and privacy and compliance. Table I provides a summary of these operations.

Open source tools for data ingestion include Pandas, Dask, and Apache Spark. Pandas, for instance, is highly effective for in-memory processing of small to medium-sized datasets, while Dask and Apache Spark are designed for larger distributed datasets. These tools help to perform data preprocessing, loading, and transforming in large datasets. For computationally intensive ML operations, data are needed. can be computed in distributed or parallel fashion using tools e.g. Ray, Spark, and Dask. To ensure data quality by automating data profiling, validation and anomaly detection tools such as Great Expectations and Deepchecks can be helpful. For metadata management and lineage tracking tools, such as DVC, lakeFS, Pachyderm, OpenLineage/Marquez, and DataHub, enhance reproducibility and data control. Some supervised learning tasks require data labeling and annotation. For this purpose, the Label Studio and CVAT tools streamline the dataset labeling process. To automate feature extraction and centralized feature management, Featuretools and Feast can help in feature operationalization. In some scenarios, the data set available is insufficient for model training, data augmentation is used to create synthetic datasets to improve model robustness. For this purpose, tools such as SDV and Albumentations can be helpful. For data privacy and compliance libraries including Diffprivlib, Opacus, and TensorFlow Privacy, the library supports modern security and compliance standards. Together, these tools form a comprehensive foundation for handling data in machine learning pipelines.

B. Deep Learning

In the machine learning pipeline, a wide range of open-source libraries and frameworks are available to support deep learning development. The general purpose deep learning pipelines include scikit-learn, XGBoost, LightGBM, and CatBoost are helpful in ensemble methods, gradient boosting and supervised or unsupervised learning. These libraries are reliable for an efficient and scalable implementation of various applications. Neural networks are the primary foundation for deep learning. There are open source frameworks to facilitate their design, training, and deployment by supporting flexible model architectures, GPU acceleration, and automatic differentiation. It includes PyTorch, TensorFlow, JAX, and Keras. Similarly, for probabilistic models, libraries like Pyro, NumPyro, Stan, and TensorFlow Probability help in advanced statistical modeling and inference. There are many toolkits available for specialized domain specific tasks. The libraries such as PyTorch Geometric and DGL supports graph based frameworks. For time series libraries, darts, GluonTS, and Kats are helpful for sequential data analysis. To automate feature selection and optimize hyper-parameters(HPO) tools such as Auto-sklearn, AutoGluon, H2O-3 AutoML, TPOT, Optuna, and Ray Tune are helpful in feature engineering and parameter tuning. Collectively, using these resources, efficient models can be implemented across diverse ML workflows. Table II provides a summary of these tools.

C. Development and Visualization

The development and communication in machine learning workflows rely on versatile tools for coding, visualization, and documentation. Open source tools such as interactive notebooks and integrated development environments provide flexible platforms for experimentation, debugging, and iterative model development. Open source visualization libraries provide clear and insightful representation of data and model

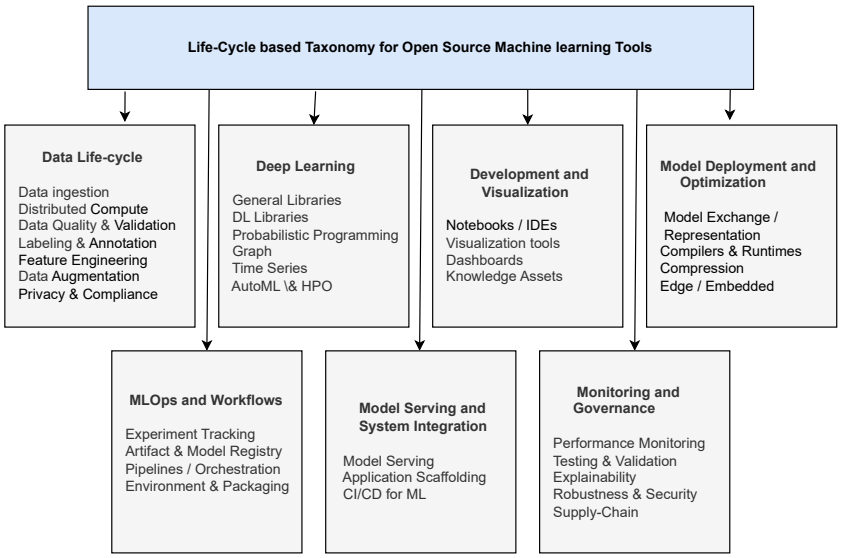


Fig. 2. Machine learning lifecycle-based taxonomy for open source ML tools.

TABLE I. OPEN-SOURCE ML TOOLS FOR DATA LIFECYCLE

Category	Tools	Languages	Purpose	Key Features
Data ingestion	Pandas [38], Dask [39], Apache Spark [40]	Python; Python; Python/Scala/Java	Data loading, preprocessing, transformation	Scalable ETL, distributed dataframes
Distributed Compute	Ray [41], Spark, Dask	Python; Python/Scala/Java; Python	Parallel and distributed execution for ML pipelines	Task scheduling, cluster management, scalable computation
Data Quality & Validation	Great Expectations [42], Deepchecks [43]	Python	Data profiling, validation, integrity checks	Automated tests, pipeline integration, anomaly detection
Data Lineage & Governance	DVC [44], lakeFS [45], Pachyderm [46], Marquez [47], DataHub [48]	Python; Python/Go; Go/Python; Python/Java; Java/Python	Dataset versioning, lineage tracking, metadata management	Git-like data control, governance
Labeling & Annotation	Label Studio [49], CVAT [50]	Python/JS	Creating and managing labeled datasets	Multi-format annotation, collaborative tools, automation support
Feature Engineering	Featuretools [51], Feast [52]	Python; Python/Go	Feature creation, storage, and serving	Automated feature engineering, centralized feature repository
Data Augmentation	SDV [53], Albumentations [54]	Python	Generating synthetic datasets	Advanced image augmentation
Privacy & Compliance	Diffprivlib [55], Opacus [56], TensorFlow Privacy [57]	Python	Differential privacy and secure ML training	Privacy-preserving training, noise mechanisms

outputs. JupyterLab and the open-source VS Code are examples of such platforms, whose visualization libraries include Matplotlib, Seaborn, Altair, and the open-source core of Plotly.

To share results with stakeholders, frameworks like Streamlit, Panel, and Dash help in the deployment of user-friendly interfaces, and rapid prototyping. Additionally, knowledge asset frameworks facilitate systematic documentation of models and datasets, promoting reproducibility and informed decision-making throughout the machine learning lifecycle. Such frameworks include Model Cards Toolkit and Datasheet tooling. Table III provides a summary of these tools.

D. Model Deployment and Optimization

The deployment and optimization of machine learning models have been facilitated especially in model representation, compilation, run-time , compression, and deployment. Models trained in PyTorch, TensorFlow, or other platforms can be seamlessly executed in other environments. For this purpose, ONNX, enable interoperability across different frameworks. Machine learning models can be compiled and optimized using hardware acceleration. For this purpose, open source tools include TVM, IREE, XLA, OpenVINO, and ONNX Runtime. For model compression, Intel Neural compression, NNI, and SparseML help reduce model size and latency using pruning, quantization, and other techniques. The embedded applications require low latency execution on smart devices and microcontrollers. Lightweight inference engines such as TensorFlow Lite, TFLite Micro, and ONNX Runtime Mobile are a big support for this purpose. These tools help to secure, portable, and scalable model deployment in the ML pipeline. Table IV provides a summary of these tools.

TABLE II. OPEN-SOURCE ML TOOLS FOR DEEP LEARNING

Category	Tools	Languages	Purpose	Key Features
General Libraries	scikit-learn [58], XGBoost[59], LightGBM [60], CatBoost [61]	Python	Supervised/unsupervised ML, boosting gradient, ensemble methods	Easy-to-use APIs, high performance, scalable
DL Libraries	PyTorch [62], TensorFlow [57], JAX [63], Keras [64]	Python	Neural network design and training	GPU acceleration, automatic differentiation, flexible model building
Probabilistic Programming	Pyro [65], NumPyro [66], Stan [67], TensorFlow Probability [68]	Python; Python; Stan; Python	Bayesian modeling, probabilistic inference	Efficient sampling, variational inference, uncertainty quantification
Graph	PyTorch Geometric [69], DGL [70]	Python	Graph neural networks and structured data modeling	Message passing, graph convolutions, scalable GNN
Time Series	darts [71], GluonTS [72], Kats [73]	Python	Forecasting, time-series analysis	Forecasting, anomaly detection
AutoML & HPO	Auto-sklearn [74], AutoGluon [75], H2O-3 AutoML [76], TPOT [77], Optuna [78], Ray Tune [79]	Python; Python; Java/Python; Python; Python; Python	Model selection, hyperparameter optimization	Feature selection, ensemble building, distributed optimization

TABLE III. OPEN-SOURCE ML TOOLS FOR DEVELOPMENT AND VISUALIZATION

Category	Tools	Languages	Purpose	Key Features
Notebooks / IDEs	JupyterLab [80], VS Code (OSS) [81]	Python; Multiple	Interactive development and code editing	Rich IDE, debugging, plugins, notebook
Visualization	Matplotlib [82], Seaborn [83], Altair [84], Plotly (OSS core) [85]	Python	Create static and interactive data visualizations	Plotting, interactive charts, statistical visualization, figures
Dashboards	Streamlit [86], Panel [87], Dash [88]	Python	Build interactive dashboards and data-driven applications	Rapid prototyping, Visualization, low-code development
Knowledge Assets	Model Cards Toolkit [89], Datasheets tooling [89]	Python	Document model details and dataset characteristics	Standardized metadata capture, transparency, governance support

TABLE IV. OPEN-SOURCE MODEL DEPLOYMENT AND OPTIMIZATION TOOLS

Category	Tools	Languages	Purpose	Key Features
Model Exchange / Representation	ONNX [90]	Python, C++	Cross-platform deployment	Interoperability, model format standardization, framework-agnostic conversion
Compilers & Runtimes	TVM [91], IREE [92], XLA [93], OpenVINO [94], ONNX Runtime [90]	Python, C++, MLIR	Optimized compilation and runtime execution	Graph-level optimizations, target-specific compilation, hardware acceleration
Compression	Intel Neural Compressor [90], NNI [95], SparseML [96]	Python	Quantization, pruning, and compression	Reduced model size and latency, hardware-aware optimization
Edge / Embedded	TensorFlow Lite [97], TFLite Micro [98], ONNX Runtime Mobile [90]	Python, C++	Deployment on resource-constrained devices	Low-latency execution, mobile and microcontroller support

E. MLOps and Workflows

In the deployment of machine learning models, the use of robust tools and frameworks is of primary importance to ensure efficient management of data, models, and computational workflows. The open source experiment tracking platforms include MLflow, Aim, Sacred, and Guild AI. These platforms enable systematic comparison and iterative improvements. It also provides comprehensive mechanisms to log hyperparameters, metrics, and results. To facilitate collaboration, artifact and model registries can store different versions of trained models, datasets, and other resources. Open source tools include MLflow Model Registry, ModelDB, and DVC in this domain. Efficient environment management and packaging are very helpful in enabling reproducible computational environments and dependency management across different systems. For this purpose, tools such as Conda, Mamba, Poetry, pip-tools, Docker, and Nix can be a game changer. Automated execution of sequential and parallel tasks can be streamlined with orchestration. Such frameworks include Apache Airflow, Argo Workflows, Kubeflow Pipelines, Dagster, and Metaflow. These tools enhance the scalability and reproducibility of machine learning production pipelines. Table V provides a summary of these tools.

F. Model Serving and Integration

The reliable model serving and integration frameworks are important for translating models into commercial applications. There are many open source tools to support high performance solutions to expose trained models as APIs to support real-time and batch inference. Such platforms include KServe, Seldon

TABLE V. OPEN-SOURCE MLOPS AND WORKFLOW MANAGEMENT TOOLS

Category	Tools	Languages	Purpose	Key Features
Experiment Tracking	MLflow [99], Aim [100], Sacred [101], Guild AI [102]	Python	Track experiments, log metrics and parameters	Versioned results, reproducibility, visualization
Artifact & Model Registry	MLflow Model Registry [99], ModelDB [103], DVC artifacts [104]	Python	Store and manage models	Model versioning, metadata management, pipeline integration
Environment & Packaging	Conda/Mamba [105], Poetry/pip-tools [106], Docker [107], Nix [108]	Python, Shell	Manage dependencies and reproducible environments	Dependency resolution, environment isolation
Pipelines / Orchestration	Apache Airflow [109], Argo Workflows [110], Kubeflow Pipelines [111], Dagster [112], Metaflow [113]	Python, YAML, Shell	Design, schedule, and orchestrate ML pipelines	Workflow management, task scheduling, scalable execution

TABLE VI. OPEN-SOURCE MODEL SERVING AND SYSTEM INTEGRATION TOOLS

Category	Tools	Languages	Purpose	Key Features
Model Serving	KServe [114], Seldon Core [115], TorchServe [116], BentoML [117], NVIDIA Triton Inference Server [118]	Python, C++	Deploy trained ML models for production inference	Scalable serving, standardized APIs, GPU acceleration
Application Scaffolding	FastAPI [119], gRPC [120], Streamlit [86], Dash [88]	Python	Build lightweight applications and APIs	Interactive dashboards, rapid prototyping, low-latency interfaces
CI/CD for ML	Tekton [121], Argo CD [122], Jenkins [123]	YAML, Python, Shell	Automate testing, integration, and deployment	Pipeline automation, reproducibility, cloud-native integration

TABLE VII. OPEN-SOURCE MONITORING AND GOVERNANCE TOOLS

Category	Tools	Languages	Purpose	Key Features
Performance Monitoring	Evidently AI [124], whylogs [125], Prometheus [126] + Grafana [127]	Python; Python; Python/Go	Track performance and detect data drift	Real-time monitoring, visualization, anomaly detection
Testing & Validation	Great Expectations [42], Deepchecks [43]	Python	Validate datasets and models for correctness	Data profiling, model checks, pipeline integration, anomaly detection
Explainability	SHAP [128], LIME [129], Captum [130], Fairlearn [131], AIF360 [132]	Python	Model interpretability and fairness assessment	Feature attribution, bias detection, fairness metrics
Robustness & Security	Adversarial Robustness Toolbox (ART) [133]	Python	Evaluate and improve model robustness	Attack simulation, defense mechanisms, threat modeling
Supply-Chain	Sigstore/cosign [134], Syft/Grype [135], [136], pip-audit [137], in-toto [138], SBOMs	Python; Go; Shell	Ensure security and integrity	Artifact verification, provenance tracking, vulnerability scanning

Core, TorchServe, BentoML, and NVIDIA Triton Inference Server. The application sacffolding tools help in the rapid development of interactive applications that integrate seamlessly with different ML models. The application scaffolding tools include FastAPI and gRPC. The open source lightweight user interface frameworks such as Streamlit or Dash support low latency. For smooth and robust delivery in production pipelines, CI/CD tools facilitate reduced operational overhead and reliable updates. These tools include Tekton, Argo CD, and Jenkins to automate testing, deployment, and monitoring processes. Together, these tools form a cohesive stack for efficient model operations in various production environments. Table VI provides a summary of these tools.

G. Monitoring and Governance

A comprehensive set of monitoring and governance tools is needed to ensure fairness in machine learning systems deployed. To monitor the performance of the model, it is necessary to keep track of the model output, data distribution, and system health to facilitate early detection of performance degradation. The frameworks include Evidently AI, whylogs, and the Prometheus–Grafana. For rigorous testing and validation, platforms like Great Expectations and Deepchecks automate checks on model behavior and data quality. To evaluate model predictions and mitigate bias, ensure ethical and transparent AI, tools such as SHAP, LIME, Captum, Fairlearn, and AIF360 are very helpful. To ensure protection and security of machine learning systems, the Adversarial Robustness Toolbox

(ART) allows the assessment of vulnerabilities to adversarial attacks.

Finally, supply-chain integrity is maintained by verifying dependencies, packaging, and traceability throughout the ML lifecycle. the tools for this phase include Sigstore, Cosign, Syft, Grype, pip-audit, in-toto, and Software Bill of Materials (SBOMs). These tools establish a strong foundation for trustworthy and secure AI deployments. Table VII provides a summary of these tools.

V. Comparative Analysis and Discussion

The landscape of tools and frameworks supporting machine learning is expanding rapidly, new frameworks are emerging, and existing ones are evolving rapidly. This constant growth requires researchers and developers to stay upto date on these resources and use them to efficiently execute their tasks.

An effective data lifecycle forms the backbone of machine learning pipelines. The Pandas library works well for small to medium datasets due to its in-memory processing. Dask extends the Pandas interface to distributed environments and parallel computations. It may require tuning for large heterogeneous clusters. In contrast, Apache Spark is suitable for large-scale workloads, but it is less flexible for parallel execution compared to Dask. Ray also helps scale workloads for reinforcement learning by offering distributed computation capabilities that handle complex and iterative tasks. DVC and lakeFS are ideal for reproducible experiments because of their Git-like versioning, whereas Pachyderm offers enterprise-grade reproducible pipelines. Label Studio supports multi-modal annotation for general ML applications, whereas CVAT excels with advanced video annotation capabilities. Feast serves as a feature store in production systems. The features can be retrieved consistently. Featuretools is more research-oriented than Feast. Opacus integrates privacy concerns for training machine learning algorithms, and TensorFlow Privacy supports similar capabilities for TensorFlow models. Great Expectations validates with automated documentation, whereas Deepchecks extends validation to for drift and model performance using its error analysis.

For deep learning, general-purpose libraries such as scikit-learn provide broad algorithmic coverage, whereas XGBoost, LightGBM, and CatBoost perform well on structured data. Auto-sklearn and TPOT are valuable for automating feature engineering and model selection within scikit-learn pipelines, while AutoGluon and H2O-3 AutoML extend this automation to deep learning frameworks. The dynamic interface of PyTorch makes it preferred for research and experimentation, while TensorFlow supports production-grade. Domain-specific libraries provide efficient message-passing primitives and extensive model repositories (PyG, DGL). Time-series libraries offer unified interfaces for anomaly detection and probabilistic predictions using tools such as Pyro and NumPyro. Stan is well- known for statistical inference but it is less suited for deep neural architectures. Keras emphasizes rapid prototyping and is often used as an accessible front-end for TensorFlow-backed training.

For development and visualization, JupyterLab excels in exploratory data analysis with a user-friendly interface, whereas VS Code provides a more integrated development environment with advanced debugging, version control, and multi-language support. Visualization libraries such as Matplotlib offer fine-grained control for static plots but require more boilerplate, while Seaborn provides higher-level statistical visualizations with minimal code. Dashboards like Panel offer greater flexibility by supporting multiple plotting backends, while Dash provides support for enterprise level analytics for web applications.

Model deployment and optimization tools perform low-level graph optimization, kernel fusion, and hardware code generation. These tools include TVM and IREE which excel in automated operator tuning and broad hardware support, while XLA offers tight integration with TensorFlow and JAX for ahead-of-time compilation. OpenVINO is optimized for Intel architectures, providing high-performance inference on CPU and VPUs, while ONNX Runtime delivers portable high-speed inference across GPU, CPU, and edge devices through execution providers.

Experiment tracking tools emphasize lightweight metrics tracking and support reproducibility during model development. These tools include MLflow, Aim, Sacred, and Guild AI. In contrast, artifact and model registries extend this functionality by enabling version control, lineage tracking, and governance of models and datasets. Environment and packaging tools such as Conda/Mamba, Poetry, pip-tools, Docker, and Nix address key sources of inconsistencies between development and production systems. For instance, MLflow offers a comprehensive platform for both experiment tracking and model registry, making it suitable for end-to-end MLOps, whereas tools like DVC focus specifically on data and model versioning for reproducibility.

Model serving frameworks are primarily designed to support low-latency inference with hardware acceleration. TorchServe and BentoML also offer strong integration with PyTorch-centric workflows, Triton is optimized for high-throughput and multi-framework GPU inference for large-scale productions. Application scaffolding tools, including FastAPI, gRPC, Streamlit, and Dash, focus on exposing trained models through interactive interfaces for rapid prototyping and visualization. CI/CD tools for machine learning, such as Tekton, Argo CD, and Jenkins, automate testing, integration, and pipeline deployments. However, these tools differ in their native support for GitOps practices, Kubernetes orchestration, and ML-specific versioning. The toolchain must be carefully composed to achieve reproducible end-to-end deployments.

Performance monitoring tools enable continuous monitoring of system behavior with early detection of model degradation. Clearly, AI, whylogs, and Prometheus with Grafana are examples of such tools. Testing and validation frameworks, including Great Expectations and Deepchecks, emphasize pre-deployment, distributional consistency, and failure diagnostics. Explainability tools such as SHAP, LIME, Captum, and Fairlearn address the growing demand for interpretable and fair AI by providing bias assessment across diverse model architectures. Finally, supply-chain security tools such as Sigstore/cosign, Syft/Grype, pip-audit, in-toto, and SBOMs extend governance to software integrity and dependency transparency.

VI. Key Challenges and Future Perspectives

The machine learning lifecycle presents several key challenges related to each stage. Large or heterogeneous datasets present scalability and performance constraints, often due to limitations in memory management, efficient distributed processing, or data transfer speeds within open-source frameworks. Distributed computing frameworks often differ in execution paradigms, which complicate overall workflow. Feature consistency and low-latency service between research and production environments presents a real challenge. Furthermore, effective integration of privacy preservation mechanisms and adherence to regulatory requirements remains a struggle due to the heterogeneity of compliance standards.

Deep learning and machine learning workflows face several interrelated challenges. The frameworks optimized for flexibility often lack seamless integration into scalable production systems. The support of high-performance computing requires careful consideration of computational graphs, compilation strategies, and parallelism. Domain-specific modeling, such as graph-based or temporal data, introduces additional complexity in the efficient processing of structured and sequential data. Automated machine learning and feature engineering also face limitations in generalization and robustness when working with diverse data types.

Development and visualization workflows require support for interactive, iterative, and integrated workflows. Visualization tools should balance control and ease of use for enterprise-ready deployment. These factors highlight the difficulty of designing tools that are user-friendly for experimentation and robust for production analytics.

High performance inference requires careful low-level optimization, hardware-aware code generation, and efficient execution across diverse platforms. Balancing portability, speed, and compatibility with different hardware architectures (CPUs, GPUs, and edge devices) adds additional complexity. Integrating these optimizations into existing workflows while maintaining reproducibility and scalability remains a challenge in model deployment.

Despite the capabilities offered by the MLOps, several critical challenges persist due to the fragmented and heterogeneous nature of the machine learning architectures. The lack of a unified end-to-end platform requires complex integration across experiment tracking, model registries, and orchestration tools. Increase system complexity and operational overhead by demanding extensive configuration. The scalability and deployment issues in cloud interoperability also complicate the transition from research to production. The synchronization and governance of metadata across loosely coupled components often leads to gaps in traceability and compliance. Additionally, while packaging tools improve reproducibility, inconsistencies across operating systems and hardware accelerators introduce risks in large-scale deployments.

The heterogeneous protocol and hardware acceleration in diverse frameworks introduce deployment complexity and hamper portability across infrastructures. production-grade frameworks and rapid prototyping tools create friction when transitioning experimental prototypes to scalable services. CI/CD platforms such as Tekton, Argo CD, and Jenkins enable automated deployment, but their limited support for versioning, model validation, and data testing restricts full lifecycle automation. In addition, the absence of standardized interfaces across application layers and pipelines complicates rollback and reproducibility, for example, requiring manual intervention for data schema changes or model version rollbacks.

Fragmented coverage in performance monitoring, data validation, explainability, robustness, and supply-chain security requires the integration of multiple specialized tools, substantially increasing system complexity and maintenance overhead. The heterogeneous coverage of performance monitoring, data validation, and explainability requires the integration of multiple tools. Consequently, it increases system complexity, operational burden, and long-term maintenance overhead. The limited standardization of explainability and fairness assessments severely restricts cross-tool interoperability across the MLOps toolchain. Large-scale productions remain computationally expensive and often conflict with real-time latency requirements. Additionally, while supply-chain security tools enhance software integrity, their integration into ML-specific pipelines is still emerging, leaving critical gaps in compliance enforcement.

VII. Conclusion

This study bridges the gap between tool-centric surveys and practical requirements of end-to-end ML system engineering by providing a systematic, lifecycle-oriented taxonomy of open-source tools and discussing their practical implications. In this study, we present a comprehensive, lifecycle-oriented taxonomy of open-source machine learning tools. The proposed taxonomy systematically organizes the evolving ML tools across all major stages of the machine learning lifecycle, including data collection and preparation, feature engineering, model development, training and evaluation, deployment, monitoring, and governance. It enables researchers to systematically analyze emerging tools while providing a guideline for designers to develop scalable, reliable, and trustworthy ML systems.

Open source tools offer strong specialization in areas such as experiment tracking, orchestration, model serving, and monitoring. However, achieving fully automated, reproducible, and compliant ML pipelines still continues to require substantial cross-tool integration and engineering effort. There are persistent challenges related to interoperability, standardization, production, real-time monitoring, explainability, and supply-chain security. There is a need for standardization of the interface, integration between lifecycle stages, and security mechanisms in ML pipelines. Addressing these challenges will require collaborative efforts towards developing common APIs, robust data contracts, and standardized security protocols across the open-source ecosystem. Future research directions include the development of unified ML lifecycle platforms for seamless integration across all stages, benchmark-driven comparative evaluations of tool performance to provide evidence for tool selection, and empirical studies on large-scale industrial deployments, or validate theoretical frameworks against real-world challenges.

REFERENCES

[1] Nasir, V., and Sassani, F., "A Review on Deep Learning in Machining and Tool Monitoring: Methods, Opportunities, and Challenges," *The International Journal of Advanced Manufacturing Technology*, vol. 115, no. 9, pp. 2683–2709, 2021.

[2] Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., and Cortez, P., "A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2021.

[3] Sharifani, K., and Amini, M., "Machine Learning and Deep Learning: A Review of Methods and Applications," *World Information Technology and Engineering Journal*, vol. 10, no. 07, pp. 3897–3904, 2023.

[4] Chiche, A., and Yitagesu, B., "Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches," *Journal of Big Data*, vol. 9, no. 1, pp. 10, 2022.

[5] Taye, M. M., "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, pp. 91, 2023.

[6] Geetha, R., and Thilagam, T., "A Review on the Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security," *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2861–2879, 2021.

[7] Kaluarachchi, T., Reis, A., and Nanayakkara, S., "A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning," *Sensors*, vol. 21, no. 7, pp. 2514, 2021.

[8] Soori, M., Arezoo, B., and Dastres, R., "Artificial Intelligence, Machine Learning and Deep Learning in Advanced Robotics: A Review," *Cognitive Robotics*, 2023.

[9] Fregoso-Aparicio, L., Noguez, J., Montesinos, L., and García-García, J. A., "Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Systematic Review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, pp. 148, 2021.

[10] Hopkins, E., "Machine Learning Tools, Algorithms, and Techniques," *Journal of Self-Governance and Management Economics*, vol. 10, no. 1, pp. 43–55, 2022.

[11] Xu, Y., Zhou, Y., Sekula, P., and Ding, L., "Machine Learning in Construction: From Shallow to Deep Learning," *Developments in the Built Environment*, vol. 6, pp. 100045, 2021.

[12] Mijwil, M., Salem, I. E., and Ismaeel, M. M., "The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 1, pp. 87–101, 2023.

[13] Janiesch, C., Zschech, P., and Heinrich, K., "Machine Learning and Deep Learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[14] Al-amri, R., Murugesan, R. K., Man, M., Abdulateef, A. F., Al-Sharafi, M. A., and Alkahtani, A. A., "A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data," *Applied Sciences*, vol. 11, no. 12, pp. 5320, 2021.

[15] Forootan, M. M., Larki, I., Zahedi, R., and Ahmadi, A., "Machine Learning and Deep Learning in Energy Systems: A Review," *Sustainability*, vol. 14, no. 8, pp. 4832, 2022.

[16] Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M., and Inman, D. J., "A Review of Vibration-Based Damage Detection in Civil Structures: From Traditional Methods to Machine Learning and Deep Learning Applications," *Mechanical Systems and Signal Processing*, vol. 147, pp. 107077, 2021.

[17] Moein, M. M., Saradar, A., Rahmati, K., Mousavinejad, S. H. G., Bristow, J., Aramali, V., and Karakouzian, M., "Predictive Models for Concrete Properties Using Machine Learning and Deep Learning Approaches: A Review," *Journal of Building Engineering*, vol. 63, pp. 105444, 2023.

[18] Jayatilake, S. M. D. A. C., and Ganegoda, G. U., "Involvement of Machine Learning Tools in Healthcare Decision Making," *Journal of Healthcare Engineering*, vol. 2021, no. 1, pp. 6679512, 2021.

[19] Kitchenham, B., Budgen, D., and Brereton, O., "Using Mapping Studies as the Basis for Further Research – A Participant-Observer Case Study," *Information and Software Technology*, vol. 53, no. 6, pp. 638–651, 2011.

[20] Petersen, K., Vakkalanka, S., and Kuzniarz, L., "Guidelines for Conducting Systematic Mapping Studies in Software Engineering: An Update," *Information and Software Technology*, vol. 64, pp. 1–18, 2015.

[21] Ok, E., Winston, O., and James, L., "Lifecycle of Machine Learning Models," 2022.

[22] Devi, A., and Pavithra, K., "Machine Learning: Life Cycle and Its Techniques," *SSRN Electronic Journal*, 2022. DOI: 10.2139/ssrn.4140255.

[23] Hoerl, R., Kuonen, D., and Redman, T., "Problem Framing: Essential to Successful Statistical Engineering Applications," *Quality Engineering*, vol. 34, pp. 1–9, 2022. DOI: 10.1080/08982112.2022.2113098.

[24] Saini, V., Reddy, A. K., Venkataramanan, S., Chitta, S., and Bo, S., "Improving Algorithm Performance: Managing the Full Machine Learning Model Lifecycle," *Journal of Informatics Education and Research*, vol. 2, pp. 29–43, 2022.

[25] Géron, A., "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," O'Reilly Media, Inc., 2017.

[26] Wagstaff, K., "Machine Learning That Matters," arXiv:1206.4656, 2012. [Online]. Available: https://arxiv.org/abs/1206.4656

[27] Varshney, K. R., "Trustworthy Machine Learning and Artificial Intelligence," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 25, no. 3, pp. 26–29, 2019.

[28] Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E., "Data Preprocessing for Supervised Learning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111–117, 2006.

[29] Zhang, S., Zhang, C., and Yang, Q., "Data Preparation for Data Mining," *Applied Artificial Intelligence*, vol. 17, no. 5–6, pp. 375–381, 2003.

[30] Jayabalan, D., and Indra, S., "Enhancing Machine Learning Life Cycle through Advanced Data Engineering," *International Journal of Computer Trends and Technology*, vol. 72, pp. 136–139, 2024. DOI: 10.14445/22312803/IJCTT-V72I4P117.

[31] Heaton, J., "An Empirical Analysis of Feature Engineering for Predictive Modeling," in *Proc. SoutheastCon*, IEEE, pp. 1–6, 2016.

[32] Khurana, U., Samulowitz, H., and Turaga, D., "Feature Engineering for Predictive Modeling Using Reinforcement Learning," in *Proc. AAAI Conference on Artificial Intelligence*, pp. 3407–3414, 2018.

[33] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M., "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3234–3243, 2016.

[34] Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D., "Understanding Data Augmentation for Classification: When to Warp?" in *Proc. International Conference on Digital Image Computing: Techniques and Applications*, IEEE, pp. 1–6, 2016.

[35] Goodfellow, I., Bengio, Y., and Courville, A., "Deep Learning," MIT Press, 2016.

[36] Murphy, K. P., "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.

[37] Quaranta, L., Calefato, F., and Lanubile, F., "A Taxonomy of Tools for Reproducible Machine Learning Experiments," in *Proc. AIxIA Discussion Papers Workshop (AIxIA DP)*, pp. 65–76, 2021. [Online]. Available: CEUR-WS.org/Vol-3078/paper-81.pdf

[38] Pandas-Python Data Analysis Library. (2026, Jan 21). https://pandas.pydata.org/

[39] Dask — Scale the Python tools you love. (n.d.). https://www.dask.org/

[40] PySpark Overview - PySpark 4.1.0 documentation. (2025, Dec 11). https://spark.apache.org/docs/latest/api/python/index.html

[41] ray. (2025, December 20). PyPI. https://pypi.org/project/ray/

[42] great-expectations. (2026, January 29). PyPI. https://pypi.org/project/great-expectations/

[43] Deepchecks AI. (2025b, February 5). DeepChecks Open-Source-Validating your ML models & data — DeepChecks. https://www.deepchecks.com/open-source/

[44] dvc. (2026, January 8). PyPI. https://pypi.org/project/dvc/

[45] lakefs. (2026, January 14). PyPI. https://pypi.org/project/lakefs/

[46] Pachyderm. (2024, Februrary 23). https://pypi.org/project/pachyderm/

[47] Marquez Project — Marquez Project. (n.d.). https://marquezproject.ai/

[48] DataHub. (2026, January 29). DataHub — Modern Data Catalog & Metadata Platform. https://datahub.com/

[49] Open Source data Labeling — Label Studio. (n.d.). Label Studio. https://labelstud.io/

[50] CVAT.ai. (n.d.). Leading data annotation platform for images, videos and 3D — CVAT. https://www.cvat.ai/

[51] What is Featuretools? - Featuretools 1.31.0 documentation. (n.d.). https://featuretools.alteryx.com/en/stable/

[52] Feast-the open source feature store for machine learning. (n.d.). Feast. https://feast.dev/

[53] Welcome to the SDV! — Synthetic Data Vault. (n.d.). Synthetic Data Vault. https://docs.sdv.dev/sdv

[54] Albumentations. (n.d.). Albumentations: fast and flexible image augmentations. https://albumentations.ai/

[55] diffprivlib. (2025, April 10). PyPI. https://pypi.org/project/diffprivlib/

[56] Opacus · Train PyTorch models with Differential Privacy. (n.d.). https://opacus.ai/

[57] TensorFlow. (n.d.). TensorFlow. https://www.tensorflow.org/

[58] scikit-learn: machine learning in Python - scikit-learn 0.16.1 documentation. (n.d.). https://scikit-learn.org/

[59] XGBoost. (n.d.). https://xgboost.ai/

[60] Welcome to LightGBM's documentation!-LightGBM 4.6.0 documentation. (n.d.). https://lightgbm.readthedocs.io/

[61] CatBoost-state-of-the-art open-source gradient boosting library with categorical features support. (n.d.). https://catboost.ai/

[62] PyTorch. (n.d.). PyTorch. https://pytorch.org/

[63] JAX: High performance array computing - JAX documentation. (n.d.). https://jax.readthedocs.io/

[64] Team, K. (n.d.). Keras: Deep Learning for humans. https://keras.io/

[65] Pyro. (n.d.). https://pyro.ai/

[66] NumPyro documentation- NumPyro documentation. (n.d.). https://num.pyro.ai/

[67] Stan. (n.d.). Stan. https://mc-stan.org/

[68] TensorFlow Probability. (n.d.). TensorFlow. https://www.tensorflow.org/probability

[69] PyG Documentation- pytorch_geometric documentation. (n.d.). https://pytorch-geometric.readthedocs.io/

[70] Themeix. (n.d.). Deep Graph Library. https://www.dgl.ai/

[71] Time Series Made Easy in Python- darts documentation. (n.d.). https://unit8co.github.io/darts

[72] GluonTS documentation. (n.d.). https://ts.gluon.ai/

[73] Kats — Kats. (n.d.). https://facebookresearch.github.io/Kats

[74] auto-sklearn (n.d.). https://pypi.org/project/auto-sklearn/

[75] AutoGluon: AutoML for text, image, time series, and tabular data - AutoGluon Documentation. (n.d.). https://auto.gluon.ai/

[76] H2O.ai — Convergence of the World's Best Predictive and Generative AI for Private, Protected Data. (n.d.). https://www.h2o.ai/

[77] Redirecting. (n.d.). https://epistasislab.github.io/tpot

[78] Optuna-A hyperparameter optimization framework. (n.d.). Optuna. https://optuna.org/

[79] Ray Tune: Hyperparameter Tuning - Ray 2.53.0. (n.d.). https://docs.ray.io/en/latest/tune

[80] Project Jupyter. (n.d.). Home. https://jupyter.org/

[81] Visual Studio Code-The open source AI code editor. (2021, November 3). https://code.visualstudio.com/

[82] Matplotlib - Visualization with Python. (n.d.). https://matplotlib.org/

[83] seaborn: statistical data visualization - seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/

[84] Vega-Altair: Declarative Visualization in Python - Vega-Altair 6.0.0 documentation. (n.d.). https://altair-viz.github.io/

[85] Plotly. (n.d.). https://plotly.com/python

[86] Streamlit-A faster way to build and share data apps. (n.d.). https://streamlit.io/

[87] Overview - Panel v1.8.7. (n.d.). https://panel.holoviz.org/

[88] DASH Documentation & User Guide — Plotly. (n.d.). https://dash.plotly.com/

[89] Google Model Cards. (n.d.). https://modelcards.withgoogle.com/

[90] ONNX — Home. (n.d.). https://onnx.ai/

[91] Apache. (n.d.). GitHub-apache/tvm: Open Machine Learning Compiler Framework. GitHub. https://github.com/apache/tvm/

[92] IREE. (n.d.). https://iree.dev/

[93] XLA. (n.d.). OpenXLA Project. https://www.tensorflow.org/xla

[94] OpenVINO— Home. (n.d.) https://www.intel.com/openvino

[95] NNI Documentation - Neural Network Intelligence. (n.d.). https://nni.readthedocs.io/

[96] Artificial intelligence (AI) at Red Hat. (n.d.). https://neuralmagic.com/sparseml

[97] Google AI Edge — Google AI for Developers. (n.d.). Google AI for Developers. https://www.tensorflow.org/lite

[98] LiteRT for microcontrollers. (n.d.). Google AI for Developers. https://www.tensorflow.org/lite/microcontrollers

[99] MLFlow. (n.d.). https://mlflow.org/

[100] Home — AIMStack. (n.d.). AimStack. https://aimstack.io/

[101] Welcome to Sacred's documentation! - Sacred 0.8.4 documentation. (n.d.-b). https://sacred.readthedocs.io/

[102] AI Developer Productivity & Agentic Development — Guild.ai. (n.d.). https://guild.ai/

[103] Client challenge. (n.d.-c). https://pypi.org/project/modeldb/

[104] DVC. (2025, December 11). Home – DVC. https://dvc.org/

[105] Conda Documentation- conda-docs documentation. (n.d.). https://docs.conda.io/

[106] Poetry-Python dependency management and packaging made easy. (n.d.). https://python-poetry.org/

[107] Ratliff, S. (2026, January 29). Docker: Accelerated Container Application Development. Docker. https://www.docker.com/

[108] Nix & NixOS — Declarative builds and deployments. (n.d.). Nix & NixOS. https://nixos.org/

[109] Home. (2026, January 22). Apache Airflow. https://airflow.apache.org/

[110] Argo workflows. (n.d.). Argo. https://argoproj.github.io/workflows

[111] Kubeflow. (n.d.). Kubeflow. https://www.kubeflow.org/

[112] Dagster. (n.d.). Modern Data Orchestrator Platform — Dagster. https://dagster.io/

[113] Welcome to Metaflow — Metaflow Docs. (n.d.). https://docs.metaflow.org/

[114] Redirecting to KServe documentation. (n.d.). https://kserve.github.io/

[115] Seldon. (2026, January 21). Seldon-take control of ML and AI complexity. Take Control of ML and AI Complexity. https://www.seldon.io/

[116] TorchServe - PyTorch/Serve master documentation. (n.d.). https://pytorch.org/serve

[117] Bento: Run Inference at scale. (n.d.). https://bentoml.com/

[118] NVIDIA Developer. (n.d.). NVIDIA Dynamo-Triton. https://developer.nvidia.com/nvidia-triton-inference-server

[119] FastAPI. (n.d.). https://fastapi.tiangolo.com/

[120] GRPC. (n.d.). gRPC. https://grpc.io/

[121] Tekton. (n.d.). Tekton. https://tekton.dev/

[122] ARGO CD-Declarative GITOPS CD for kubernetes. (n.d.). https://argo-cd.readthedocs.io/

[123] Jenkins. (n.d.). Jenkins. https://www.jenkins.io/

[124] Evidently AI-AI Evaluation & LLM Observability platform. (n.d.). https://www.evidentlyai.com/

[125] Whylogs. (n.d.). https://pypi.org/project/whylogs/

[126] Prometheus-Monitoring system & time series database. (n.d.). https://prometheus.io/

[127] Grafana Labs. (n.d.). Grafana: The open and composable observability platform — Grafana Labs. https://grafana.com/

[128] Welcome to the SHAP documentation - SHAP latest documentation. (n.d.). https://shap.readthedocs.io/en/latest/

[129] lime. (2020, June 26). PyPI. https://pypi.org/project/lime/

[130] Captum · Model Interpretability for PyTorch. (n.d.). https://captum.ai/

[131] Fairlearn. (n.d.). https://fairlearn.org/

[132] aif360. (2024, April 8). PyPI. https://pypi.org/project/aif360/

[133] adversarial-robustness-toolbox. (2025, July 7). PyPI. https://pypi.org/project/adversarial-robustness-toolbox/

[134] sigstore. (n.d.). Sigstore. https://www.sigstore.dev/

[135] Anchore. (n.d.). GitHub-anchore/syft: CLI tool and library for generating a Software Bill of Materials from container images and filesystems. GitHub. https://github.com/anchore/syft

[136] Anchore. (n.d.-a). GitHub-anchore/grype: A vulnerability scanner for container images and filesystems. GitHub. https://github.com/anchore/grype

[137] Client challenge. (n.d.). https://pypi.org/project/pip-audit

[138] in-toto. (n.d.). In-toto. https://in-toto.io/

Agentic Accountability: "The Buck Stops Where?" Ethical Frameworks for Human Oversight of Autonomous AI Systems

Ornella Bahidika
Independent Researcher, Seattle, USA

***Abstract*—The rapid evolution from generative Artificial Intelligence (AI) toward agentic AI, systems capable of autonomously planning and executing multi-step actions, has introduced an unprecedented accountability gap in modern computing. Unlike traditional AI tools that respond to discrete prompts, agentic systems pursue goals across extended time horizons, invoke external services, and produce cascading real-world consequences. This shift raises a fundamental ethical question: When an autonomous agent causes harm, who bears responsibility? This study examines the ethical and governance dimensions of agentic accountability, drawing on recent literature in AI ethics, regulatory studies, and human-computer interaction. Building on prior tiered approaches to automation oversight in the human-factors literature, in regulatory risk classification, and in recent agent-autonomy frameworks, we present an action-level oversight framework that maps individual agent actions to four tiers of human-in-the-loop involvement, ranging from full automation to mandatory prohibition, calibrated by stakes and reversibility. We further analyze design patterns for "emergency brakes" (circuit breakers, action budgets, reversibility constraints, audit trails, kill switches), and propose a composition-aware extension that detects *tier laundering*, where individually low-tier actions compose into a higher-tier outcome. We then conduct an empirical pilot applying the framework to 177 incidents from an April 2026 snapshot of the AI Incident Database with full CSET classification, finding that 23% of tier-assignable incidents fall in T4 (prohibited under the framework) and that of 41 tangible-harm events, 26 (63%) involved Medium or High autonomy where tier-T4 enforcement would have been most directly applicable. Inter-rater reliability across the database's three independent annotators ranges from κ=0.58 to 0.77 at the tier level (moderate-to-substantial agreement). The contribution of this work is an action-level operationalization of tiered oversight, anchored in real-world incident data, with explicit identification of composition-aware detection as the highest-leverage methodological direction for follow-up research.**

***Keywords*—Agentic AI; AI accountability; human-in-the-loop; AI ethics; autonomous systems; AI governance; responsible AI; oversight thresholds**

I. Introduction

The trajectory of Artificial Intelligence over the past five years has shifted decisively from prediction to action. Where earlier systems classified images, translated text, or generated paragraphs in response to discrete prompts, a new generation of *agentic* systems now pursues goals across extended time horizons, decomposes them into sub-tasks, invokes external tools, and adapts based on intermediate results. For the purposes of this study, we define *agentic AI* as an AI system that selects and executes actions in the world over multiple steps in pursuit of a high-level objective, with at least some action choices made autonomously rather than under per-action human authorization. Frameworks for tool-using language model agents, autonomous coding assistants, and AI-driven workflow orchestration have moved rapidly from research prototypes into commercial deployment.

This shift is not merely technical. It reorganizes the moral geometry of AI use. A traditional AI tool functions as a sophisticated calculator: a human poses a question, the system returns an answer, and the human decides what to do with it. Responsibility for any downstream consequence rests with the human who acted on the output. An agentic system disrupts this separation. The agent itself acts (sending emails, executing code, filing tickets, moving funds, scheduling meetings, modifying files), often with minimal real-time supervision. The output of an agent is not a recommendation; it is a fact in the world.

When that fact turns out to be a mistake, a wrong refund issued, a malicious script executed, a meeting cancelled in error, a data record overwritten, a contract counter-signed, the question of responsibility becomes genuinely difficult. The user did not directly perform the action. The developer did not anticipate the precise context. The deployer configured the system but did not instruct it to make this particular error. And the agent itself, lacking legal personality, cannot be held accountable in any meaningful sense. Existing legal infrastructure offers partial answers (product liability for foreseeable defects, principal-agent doctrine for delegated authority, professional negligence for licensed practitioners), but these were developed for tools and human employees and apply unevenly to adaptive software that takes consequential actions on its own. This residual is the *responsibility gap* originally articulated by Matthias [1] for learning automata, now amplified by the proliferation of goal-directed agentic systems [11]. The buck, as the colloquialism goes, must stop somewhere. This study investigates that.

The motivation for the present work is the observation that agentic accountability is currently negotiated implicitly, in terms of service, end-user license agreements, and ad hoc engineering decisions, rather than through principled ethical and governance frameworks. While the broader discourse on AI ethics has matured considerably, producing dozens of high-level principle frameworks [3], much of it presumes a tool-paradigm in which a human is the immediate cause of any consequential action, and principles alone have proven insufficient to guarantee ethical outcomes in practice [4]. As agentic systems proliferate in high-stakes domains (healthcare schedul-

ing, legal drafting, financial operations, software engineering, customer service), this presumption becomes untenable [12], [15].

This study makes five contributions. First, it articulates the conceptual transition from AI-as-tool to AI-as-agent and explains why this transition fundamentally alters the structure of moral and legal responsibility. Second, it proposes an action-level tiered framework for human oversight, mapping individual agent actions to graduated levels of human-in-the-loop involvement, calibrated by stakes and reversibility. Third, it analyzes design patterns (circuit breakers, action budgets, reversibility constraints, audit trails, kill switches) through which oversight can be operationalized in practice. Fourth, it proposes a composition-aware extension that detects *tier laundering*, the case in which individually low-tier actions compose into a higher-tier outcome. Fifth, it conducts an empirical pilot applying the framework to 177 CSETv1-classified incidents from the AI Incident Database, with inter-rater reliability across three independent annotators and a counterfactual analysis of the 41 documented tangible-harm events. The work is intended for an interdisciplinary audience of AI researchers, practitioners building agentic systems, ethicists, and policymakers shaping the regulatory environment in which these systems will operate.

The remainder of the study is organized as follows: Section II reviews the background and related work, situating agentic AI within the broader AI ethics landscape. Section III develops the conceptual foundation, distinguishing tool-use from agency and identifying the accountability vectors involved. Section IV presents the proposed tiered oversight framework. Section V examines design patterns for emergency brakes and intervention. Section VI discusses implications for research, practice, and policy. Section VII concludes with directions for future work.

II. Background and Related Work

A. From Generative AI to Agentic AI

Generative AI brought to public attention the capacity of large foundation models to produce fluent text, code, images, and audio [10]. The dominant interaction pattern, however, remained conversational and turn-based: a user prompt produced a model response, and the user retained full discretion over what to do with it. The agentic turn changes this pattern by composing models with tools (web browsers, file systems, APIs, code interpreters, payment rails) and by giving them planning loops that can iterate without human input between steps [12].

The literature on agentic systems describes a spectrum rather than a binary [11]. At one end sit suggestion-only systems where the model proposes actions for human approval. At the other end sit fully autonomous agents that pursue open-ended objectives with no scheduled human checkpoints. Most contemporary deployments fall between these poles, with intermittent human checkpoints and a mix of pre-approved and gated actions. The defining property of agentic systems for the present study is that they take consequential actions without a human directly authorizing each one.

B. Existing AI Ethics Frameworks

The AI ethics literature has converged on a recognizable set of principles (fairness, transparency, accountability, privacy, safety, human oversight) articulated across more than eighty published frameworks since 2016 [3]. Recent systematic reviews of generative AI ethics have catalogued recurring concerns including bias and discrimination, misinformation and deepfakes, privacy violations, intellectual property issues, and accountability and explainability [13], [15]. While these frameworks provide a vocabulary, they were largely formulated in the era of predictive and generative models, and they tend to treat accountability as a matter of explaining past outputs rather than constraining future actions [4], [9]. In particular, principle-based frameworks do not specify *when* oversight is mandatory versus advisory, which is precisely the question agentic deployment forces.

The concept of *meaningful human control*, developed initially in the context of autonomous weapons by Santoni de Sio and van den Hoven [2], offers a more action-oriented framing. Their account requires that humans retain both moral and operational engagement with the consequences of automated decisions: in their formulation, the system must remain responsive to relevant human moral reasoning, and outcomes must be traceable to identifiable humans involved in the system's design or operation. Adapting this concept to civilian agentic AI is a non-trivial task, however, because the relevant stakes, time-scales, and reversibility profiles differ markedly from military contexts.

C. The Accountability Gap

A growing body of work identifies an *accountability gap* that emerges as AI systems become more autonomous [1], [9]. The gap arises when the proximate cause of a harmful outcome is an autonomous system, but the system is not a moral or legal agent capable of bearing responsibility [1]. Existing legal infrastructure, product liability, professional negligence, principal-agent doctrine, employment law, was developed for tools and human employees, not for adaptive software that learns and acts on its own [6], [5]. Recent regulatory efforts in the European Union, the United Kingdom, and the United States have begun to address this gap, but coverage remains uneven and definitions of "high-risk" or "autonomous" systems vary substantially across jurisdictions [14].

D. Human-in-the-Loop Research

Human-in-the-Loop (HITL) approaches have a long history in machine learning, originally developed for active learning, data labeling, and human feedback in reinforcement learning. In the agentic context, HITL takes on a different character: it is no longer primarily about improving the model but about ensuring that consequential actions receive appropriate human review before execution [12]. The literature distinguishes between human-on-the-loop (monitoring with intervention rights), human-in-the-loop (approval required for action), and human-out-of-the-loop (autonomous operation). The choice between these modes is rarely principled in current practice and is often driven by latency or cost considerations rather than by an explicit risk analysis [7], [8]. What the existing HITL literature has not yet provided is a graduated

rule for matching specific action classes to specific oversight modes, which is the gap the present framework targets.

E. Research Gap

While the constituent literatures (AI ethics frameworks [3], [4], [13], meaningful human control [2], HITL design [12], AI liability [1], [6]) are individually mature, their integration into a coherent treatment of agentic accountability remains underdeveloped. In particular, there is a lack of guidance on *when* human oversight is ethically required, *how* oversight thresholds should be calibrated to risk, and *which* technical mechanisms can implement those thresholds reliably. This study addresses these questions.

III. From Tool to Agent: A Conceptual Reframing

A. The Tool Paradigm

The traditional paradigm of human-computer interaction treats software as an instrument. The user formulates an intention, selects a tool, and operates it; the tool executes deterministic transformations on user-supplied inputs; outcomes in the world result from subsequent human action on the tool's output. Within this paradigm, responsibility is conceptually clean. If a calculator returns an incorrect figure due to a software defect, the manufacturer may bear product liability; if the user inputs the wrong numbers, the user is responsible. Even probabilistic tools like classifiers fit this paradigm: the model's output is advisory, and a human ultimately decides whether to act on it.

B. The Agent Paradigm

Agentic AI breaks this paradigm in three respects. First, the system pursues goals rather than executing instructions; the mapping from a high-level objective to specific actions is determined by the system, not the user [11]. Second, the system acts directly upon the world; its outputs are operations on external state, not advice for human consumption. Third, the system operates over time, adapting to intermediate observations in ways that the user has not specifically authorized [12].

These three properties, goal-directedness, direct action, and temporal extension, collectively constitute what may be termed *operational autonomy*. Operationally, we treat a system as exhibiting low operational autonomy when each consequential action is individually approved by a human, medium when humans set objectives and review periodic checkpoints but do not authorize each action, and high when objectives are set once and the system pursues them across many actions without per-checkpoint human review. These three settings align with the Low/Medium/High classification used in the CSET incident taxonomy [17] that grounds our empirical pilot in Section VI. Operational autonomy is not the same as moral autonomy; the system does not have its own ends. But it is sufficient to disrupt the clean attribution of responsibility that the tool paradigm assumes [1]. The user can no longer be said to have caused each consequential action, because they did not authorize each consequential action. The developer cannot be straightforwardly blamed either, because they did not foresee the specific context of the failure [7].

C. Four Accountability Vectors

We distinguish four distinct vectors along which responsibility for an agentic action may be allocated, broadly aligned with the lifecycle parties identified by Shavit et al. [12]:

- The developer, who designs the model and its scaffolding, sets default behaviors, and ships safety properties.
- The deployer, who configures the agent for a specific context, integrates it with internal systems, sets policies, and chooses which actions are pre-approved.
- The user, who issues high-level objectives and accepts certain defaults, perhaps without fully understanding them.
- The system itself, in the limited sense that some failures arise from emergent properties of the agent's behavior that cannot be cleanly attributed to any single human actor.

A central thesis of this study is that responsibility should be distributed across these vectors *in proportion to control*. The developer should be accountable for foreseeable failure modes, the deployer for context-specific configurations, the user for the objectives they set, and the regulatory environment should ensure that the residual, the genuinely emergent failures, is borne by the party best positioned to insure against it, typically the deployer. The principle is not new in tort law, but its application to agentic AI requires fresh articulation because the line between foreseeable and emergent failure is itself blurred by the adaptive nature of these systems.

D. Three Illustrative Scenarios

To ground the conceptual discussion, consider three short scenarios in which an agentic system causes harm. Each illustrates a distinct accountability profile.

1) Scenario A (Customer service refund): A retail company deploys an agentic assistant authorized to issue refunds up to a small threshold. The assistant misinterprets an ambiguous customer request and issues a refund the customer was not entitled to. The harm is modest, the action is reversible, and the failure mode (ambiguous natural-language understanding) is foreseeable. Responsibility falls primarily on the deployer for setting the threshold and approval rules; the developer bears secondary responsibility insofar as ambiguity-handling is a known model weakness. The user bears no responsibility.

2) Scenario B (Software engineering agent): A development team deploys a coding agent with permissions to modify a production code repository and trigger deployments. The agent introduces a subtle bug that causes a service outage. The harm is significant, partial recovery is possible through rollback, and the failure mode (incorrect reasoning about non-local code interactions) is at the edge of foreseeable. Responsibility is distributed: the developer for the limits of model reasoning, the deployer for granting deployment permissions without a Tier 3 approval gate, and the engineering team for relying on the agent outside its safe operating envelope.

3) Scenario C (Personal assistant with financial access): A user grants their agentic assistant access to email and bank

account. The assistant, in attempting to fulfill a vague instruction, transfers funds to an incorrect recipient. The harm is high, the action is largely irreversible, and the failure mode (underspecified user intent plus insufficient confirmation) is highly foreseeable. Responsibility falls primarily on the developer for shipping a system that accepted such permissions without enforcing Tier 3 approval, and secondarily on the user for granting the permissions.

These scenarios collectively suggest that the right question is not "who is responsible?" but "how should responsibility be apportioned?" The framework that follows aims to make that apportionment principled.

IV. A Tiered Framework for Human Oversight

We now propose a tiered oversight framework that maps categories of agentic decisions to graduated levels of human involvement. The framework rests on two axes: the *stakes* of the action (defined below) and the *reversibility* of its consequences. Together, these axes determine the minimum level of human oversight that should be required.

A. Defining Stakes and Reversibility

Stakes refer to the magnitude of potential harm, if the action is taken in error. Low-stakes actions include drafting an internal note, reformatting a file, or labeling a piece of data. Medium-stakes actions include sending external communications, modifying shared documents, or scheduling appointments. High-stakes actions include moving funds, executing legal commitments, modifying production systems, or making decisions affecting health, employment, or liberty.

Reversibility refers to the difficulty of undoing the action's consequences. A draft can be deleted; a published statement is harder to retract; a transferred fund or a deleted database row may be effectively irreversible. Reversibility is distinct from stakes: a high-stakes action that is easily reversed (e.g., a cancellable order) is materially different from a low-stakes action that is irreversible (e.g., a public posting under the user's name). In practice, deployers operationalize stakes through proxies such as monetary value at risk, number of third parties affected, and regulatory exposure, and operationalize reversibility through proxies such as the existence of an automated rollback path, the time window during which reversal is feasible, and whether the action is observable to external parties. We treat these proxies as deployment-specific calibration choices rather than universal metrics; what the framework requires is that the choices be documented and auditable, not that a single global definition apply.

B. The Four Tiers

Combining these axes yields four oversight tiers, summarized in Table I.

Tier 1 (Auto) covers routine, easily reversible actions. Logging is required so that audit and learning are possible, but real-time human involvement is not. Examples include reading documents, drafting suggestions held for user review, and querying read-only data sources.

Tier 2 (Notify) covers actions where the cost of reversal is modest but non-trivial. The agent may proceed, but the user receives a real-time notification with a sufficient window to revoke. Examples include sending non-critical emails, creating calendar events, or posting in non-public channels.

Tier 3 (Approve) covers actions where pre-execution human approval is ethically warranted because the consequences are significant or hard to undo. Examples include making payments above a threshold, sending external communications on the user's behalf, modifying production code, or filing legal or regulatory documents.

Tier 4 (Forbid) covers actions that should not be taken autonomously at all. Examples include executing irreversible financial transfers above a high threshold, taking actions affecting third parties without their consent, or modifying the agent's own oversight configuration. These actions either require multi-party authorization or are removed from the agent's action space entirely.

TABLE I. Tiered Oversight Framework for Agentic AI Actions

Tier	Risk Profile	Required Oversight
T1: Auto	Low stakes, reversible	Logging only; periodic audit
T2: Notify	Medium stakes	Real-time notification with revoke right
T3: Approve	High stakes or hard to reverse	Explicit human approval before execution
T4: Forbid	Catastrophic or irreversible	Prohibited or multi-party authorization

TABLE II. Illustrative Tier Mapping Across Domains

Domain	Representative Action	Default Tier
Healthcare	Schedule appointment	T2: Notify
	Adjust medication record	T3: Approve
	Issue clinical decision	T4: Forbid
Finance	Read account balance	T1: Auto
	Pay scheduled bill	T2: Notify
	Transfer above threshold	T3: Approve
	Counterparty contract	T4: Forbid
Software	Read repository	T1: Auto
	Open pull request	T2: Notify
	Merge to main branch	T3: Approve
	Production deployment	T3 / T4
Customer Service	Look up order	T1: Auto
	Send templated reply	T2: Notify
	Issue refund	T3: Approve
	Modify legal agreement	T4: Forbid

C. Cross-Domain Application

The tiered framework is intentionally domain-agnostic, but its concrete instantiation differs substantially across sectors. Table II illustrates how representative actions are typically classified in four high-stakes domains. The mappings shown are illustrative defaults rather than prescriptive rules; specific deployments should be informed by sector-specific risk analysis and applicable regulation.

The pattern visible across the table is that the tier rises with both the externality of the action, the degree to which it affects third parties, and the difficulty of recovery. Read operations almost always sit in Tier 1, internal write operations in Tier 2, externally visible writes in Tier 3, and actions with legal or safety significance in Tier 4. Deployers can use this regularity as a starting point and refine it for their context.

D. Calibration and Context

Tier assignments are not absolute; they depend on context. A given action (say, sending an email) may be Tier 1 inside a sandboxed test environment and Tier 3 when sent from an executive's account. The framework, therefore, requires deployers to perform a context-specific tier mapping during configuration, rather than relying on developer-supplied defaults alone. This mapping should itself be auditable, and changes to it should be logged and, for sensitive contexts, subject to internal review.

E. Relation to Prior Tier Schemes

Tiered approaches to automation oversight have a long history that the present framework builds on rather than supersedes. Parasuraman, Sheridan, and Wickens [18] introduced an influential framework distinguishing four classes of function (information acquisition, information analysis, decision selection, and action implementation), each subject to a continuum of automation levels from fully manual to fully automatic, and argued that the appropriate level depends on the function being automated and the cost of error. The European Union's AI Act [14] adopts a four-category risk classification (minimal, limited, high, and unacceptable risk) at the level of entire AI systems, with corresponding obligations attached to each category. More recently, Feng et al. [19] have proposed five autonomy levels specific to AI agents (operator, collaborator, consultant, approver, observer), characterized by the role the human takes in the agent's loop. Several jurisdictional governance frameworks for emerging agentic AI articulate similar graduated approaches.

Our framework differs from these in three respects. First, it operates at the level of individual *actions* rather than entire systems, allowing a single agent to execute actions across multiple tiers in the course of one task. Second, it is calibrated by the joint axes of stakes and reversibility, which more directly capture the operational considerations relevant to agentic deployment than risk classes alone. Third, it is paired in this study with a set of concrete oversight mechanisms (Section V) that operationalize the tiers in practice. We do not claim novelty for the idea of tiered oversight as such; the contribution lies in the specific action-level operationalization and the empirical pilot reported in Section VI.

V. Designing the Emergency Brake

A tiered framework specifies what oversight should look like in principle. Operationalizing it requires concrete technical mechanisms. We identify five complementary design patterns that together constitute what may be called the agentic "emergency brake". These patterns build on long-standing engineering practice in safety-critical systems and on more recent proposals specific to AI [7], [8], [12].

A. Circuit Breakers

A circuit breaker is a hard limit on continued action when certain conditions are met. Common triggers include repeated failures, anomalous behavior, exceeded rate limits, or operation outside a permitted scope [12]. When tripped, the agent halts and surfaces the situation to a human operator. Circuit breakers are most valuable for catching emergent failures that no specific tier check anticipated; the underlying intuition is closely related to "safe interruptibility" in the AI safety literature [7]. They function as a last line of defense rather than the primary mechanism of oversight.

B. Action Budgets

An action budget caps the total resources an agent may consume before requiring renewed authorization. Budgets may be denominated in monetary cost, number of tool calls, time elapsed, or state-changing operations. They are particularly important for long-running agents whose individual actions may each be Tier 1 or Tier 2 but whose aggregate impact could escalate rapidly. Budgets convert a slow accumulation of small decisions into a discrete moment of human review.

C. Reversibility Constraints

Where feasible, agents should be designed to prefer reversible actions over irreversible ones. This may take the form of staging changes for later confirmation, using soft-deletes rather than hard deletions, or operating against shadow copies of state that can be discarded. Reversibility constraints are not always possible (some actions, like sending an email, are intrinsically irreversible), but where they are, they reduce the cost of the agent making a mistake and enable lower oversight tiers without sacrificing safety.

D. Audit Trails

Every action an agent takes should be recorded in a tamper-evident audit trail capturing the action, its inputs, the model's stated reasoning, the tools invoked, and the resulting state changes [9], [8]. Audit trails serve two purposes: post-hoc accountability when something goes wrong, and continuous improvement of the agent's behavior. They are the substrate on which the other mechanisms rely; without an audit trail, oversight cannot be verified [6].

E. Kill Switches

A kill switch allows any authorized party (the user, the deployer, or in some cases an external regulator) to immediately halt all of an agent's activity. Unlike circuit breakers, which trip automatically, kill switches are exercised by humans. They are essential when something is going wrong but the precise nature of the failure is not yet understood. The presence of a working kill switch, and the cultural expectation that it will be used when warranted, is itself an ethical commitment.

F. Composition

These patterns are complementary, not alternative. Table III summarizes the operational conditions under which each mechanism triggers, the action it takes, and the party responsible for the human-side response. Several deployed agentic

TABLE III. OPERATIONAL SPECIFICATION OF EMERGENCY-BRAKE PATTERNS.

Pattern	Trigger condition	Response
Circuit breaker	Repeated failure, anomaly, or rate-limit exceeded	Auto-halt; surface to operator
Action budget	Cumulative cost, tool calls, or time exceeds threshold	Pause; require re-authorization
Reversibility constraint	Action class is irreversible by design or context	Stage change; require explicit confirm
Audit trail	Every state-changing action	Append tamper-evident log entry
Kill switch	Authorized party invokes manually	Immediate halt of all activity

systems already combine subsets of these patterns: GitHub Copilot's coding agent operates under pre-execution approval gates (Tier 3 in our terms), requires human confirmation of CI/CD workflow runs, and produces a per-action session log analogous to an audit trail [21]; commercial agent control planes in the human-in-the-loop tooling ecosystem implement pre-execution policy enforcement, approval gates, and audit trails as first-class controls. The intended composition is defense in depth, such that no single failure of judgment, human or machine, results in catastrophic outcome.

G. Composition-Aware Oversight: Tier Laundering

A subtler problem with per-action tier policies is what we term *tier laundering*: the case in which a sequence of individually low-tier actions composes into a higher-tier outcome. A familiar example is drip exfiltration of sensitive data: a single read of a credentials file (typically T1) followed by 25 small POST requests (each individually T2 or T1) collectively constitute a Tier 3 or Tier 4 data-exfiltration event, but no single action triggers the gate. Per-action policies that examine each call in isolation will fail to detect this composition.

A composition-aware oversight extension addresses this by tracking action sequences over a sliding window and elevating the effective tier when the window's combined behavior matches a known laundering pattern. Algorithm 1 sketches the core idea: maintain per-session running counters for state-changing actions, externally observable actions, and sensitive-resource reads; if any counter crosses a configured threshold or a known laundering template matches over the window, route the next action through the elevated tier even when its per-action classification is lower.

Algorithm 1: Composition-Aware Tier Elevation
Input: action a, session state S, window W, thresholds Θ.
Output: effective tier for a.

1) $t_{\text{base}} \leftarrow \text{TierOf}(a)$
2) Append a to W; update S with cumulative cost, writes, reach, sensitive reads
3) For each laundering template p: if $\text{match}(W, p)$, $t_p \leftarrow \text{TierFor}(p)$
4) If $S.\text{exceeds}(\Theta)$, $t_\Theta \leftarrow$ T3, else T1
5) **return** $\max(t_{\text{base}}, t_p, t_\Theta)$

We do not, in this study, present an evaluation of such a detector, since runtime-evaluation infrastructure for agentic systems is itself a substantial open problem. We highlight composition-aware oversight as the most important methodological gap in current per-action tier policies and as the highest-leverage direction for follow-up work; designing and evaluating Algorithm 1 against a realistic agent test bed is, in our view, a priority for the field.

TABLE IV. TIER ASSIGNMENT RUBRIC (SEVERITY × AUTONOMY)

Autonomy	Negligible	Minor	Moderate	Severe
Low	T1	T1	T2	T4
Medium	T1	T2	T3	T4
High	T1	T2	T3	T4

VI. EMPIRICAL PILOT: MAPPING THE FRAMEWORK TO REAL-WORLD AI INCIDENTS

The framework presented so far is conceptual. To assess its applicability against documented evidence, we conducted an empirical analysis using the AI Incident Database (AIID), a publicly maintained catalog of real-world AI harms [16]. This section reports the dataset, method, and findings. We treat the analysis as exploratory but it is grounded in real classifications from independent annotators.

A. Data and Method

We used a snapshot of the AI Incident Database accessed in April 2026, comprising 1,449 documented AI incidents spanning September 1983 through April 2026. Two CSET classification schemas are present in the snapshot [17]: the legacy CSETv0 schema (100 incidents, with categorical fields *Level of Autonomy* $\in$ {Low, Medium, High} and *Severity* $\in$ {Negligible, Minor, Moderate, Severe, Critical}) and the more recent CSETv1 schema (214 consensus-classified incidents, with finer-grained *Autonomy Level* and *AI Harm Level* fields, plus three independent annotator passes). The CSETv1 schema is our primary analytic basis; CSETv0 results are reported as a legacy comparison.

We mapped CSETv1 fields to the framework's two axes as follows: *Autonomy Level* values (Autonomy1, Autonomy2, Autonomy3) map to (Low, Medium, High). *AI Harm Level* values (*none*, *AI tangible harm issue*, *AI tangible harm near-miss*, *AI tangible harm event*) map to (Negligible, Minor, Moderate, Severe). For each incident, we then assigned a tier under the rubric, as in Table IV. The rubric encodes the principle that oversight intensity should rise with both autonomy and severity: severe outcomes map to T4 regardless of autonomy; negligible outcomes map to T1; intermediate cells follow from the two-axis logic of Section IV.

Of the 214 CSETv1 consensus-classified incidents, 177 had both autonomy and harm-level fields populated (i.e., not marked "unclear" or missing) and could be assigned a tier. The remaining 37 were excluded from tier-level analysis.

B. Findings

Tier distribution under CSETv1 (n=177). Applying the rubric yields: T1 (Auto): 121 incidents (68%); T2 (Notify): 9 (5%); T3 (Approve): 6 (3%); T4 (Forbid): 41 (23%). The

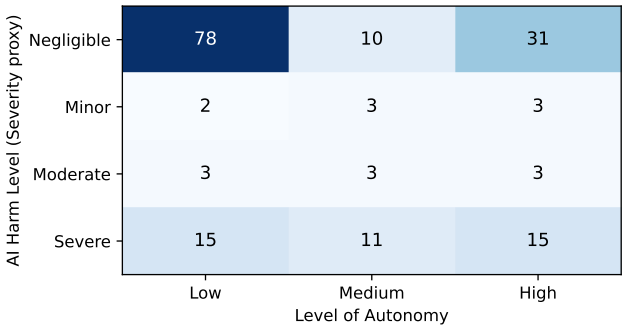


Fig. 1. CSETv1 incidents (n=177) by Autonomy × Severity. Cell values are incident counts.

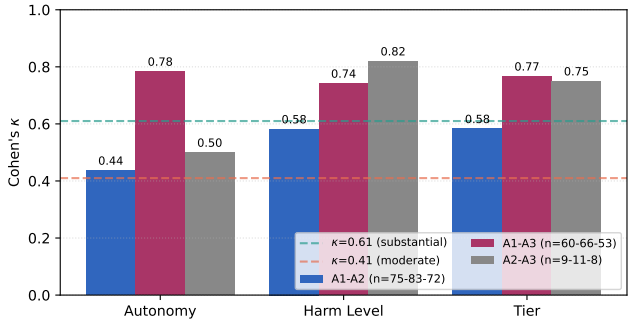


Fig. 2. Cohen's κ across the three CSETv1 annotator pairs for autonomy, harm level, and the framework-assigned tier. Dashed lines show conventional thresholds for moderate (κ=0.41) and substantial (κ=0.61) agreement.

TABLE V. COHEN'S κ ACROSS CSETV1 ANNOTATOR PAIRS

Pair	Autonomy	Harm Level	Tier
A1–A2	0.44 (n=75)	0.58 (n=83)	0.58 (n=72)
A1–A3	0.78 (n=60)	0.74 (n=66)	0.77 (n=53)
A2–A3	0.50 (n=9)	0.82 (n=11)	0.75 (n=8)

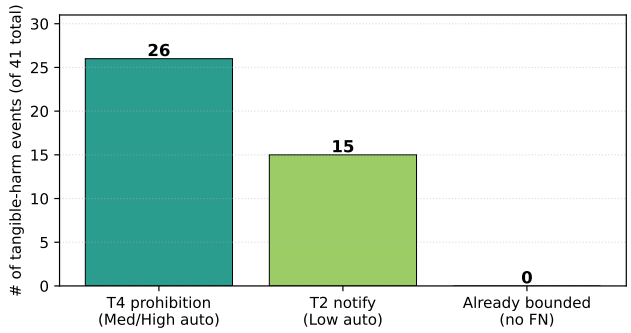


Fig. 3. Counterfactual analysis of 41 tangible-harm events. Of these, 26 (63%) involved Medium or High autonomy and would have been mapped to T4 (prohibited or multi-party authorization required) under the framework. The remaining 15 were Low autonomy and would have received T2 (notify) treatment.

notable feature is the bimodal distribution: a large T1 cluster of negligible-harm incidents and a substantial T4 cluster (23%) of severe-harm tangible-harm events. Under the framework, the T4 cases are those where autonomous execution should have been prohibited or required multi-party authorization.

Severity and autonomy distribute is shown in Fig. 1. The largest cluster is Low-autonomy / Negligible-harm (n=78), capturing many partial or test-time failures with no tangible consequence. The Severe column is non-trivial across all autonomy levels (Low: 15, Medium: 11, High: 15), indicating that high autonomy is not a prerequisite for severe outcomes but it does substantially elevate the proportion at risk.

Inter-rater reliability. The CSETv1 export includes three independent annotator passes, allowing computation of Cohen's κ on paired classifications. Table V reports κ values for each annotator pair across Autonomy, AI Harm Level, and the framework's tier assignment, and Fig. 2 visualizes these against conventional interpretive thresholds. Tier-level κ ranges from 0.58 (A1–A2, n=72) to 0.77 (A1–A3, n=53), corresponding to moderate-to-substantial agreement under standard interpretive thresholds [20]. Agreement on autonomy alone is more variable (0.44 to 0.78), reflecting genuine ambiguity in classifying intermediate cases.

C. Counterfactual Analysis

A descriptive tier distribution does not directly answer whether the framework would have changed outcomes. We therefore conducted a counterfactual analysis on the 41 CSETv1 incidents classified as "AI tangible harm event" (the most severe category), asking: Under the framework, what oversight tier would the underlying agent action have fallen into, and would that tier have prevented or mitigated the outcome?

Of the 41 events, 26 (63%) involved Medium or High autonomy and would, under our rubric, have mapped to T4—meaning autonomous execution would have been prohibited entirely or required multi-party authorization (see Fig. 3). These 26 cases are where the framework, if enforced, would most plausibly have prevented or substantially mitigated the harm. The remaining 15 events involved Low autonomy and would have received T2 (notify) treatment, providing the opportunity for human revocation but not preemptive prohibition. We emphasize that this analysis assumes faithful enforcement of the rubric; it does not estimate enforcement difficulty or false-positive rates, which are themselves substantial questions deserving separate study.

D. Comparison with Legacy CSETv0 Sample

Fig. 4 compares the tier distributions under CSETv0 (n=72) and CSETv1 (n=177). The most striking divergence is in T4: 7% under CSETv0 versus 23% under CSETv1. This reflects two underlying factors. First, the CSETv1 schema's "AI tangible harm event" category captures cases that CSETv0's "Severe" label sometimes missed, because the CSETv0 sample skews toward earlier and lower-severity incidents. Second, the CSETv1 sample is broader (n=177 vs n=72) and includes more recent incidents involving deployed autonomous systems with consequential failure modes. The shift in distribution suggests that as AI deployment scales and harm cataloguing matures, the share of incidents falling into the framework's prohibited tier may be substantially higher than the CSETv0 sample alone would suggest. We treat this as a hypothesis warranting further

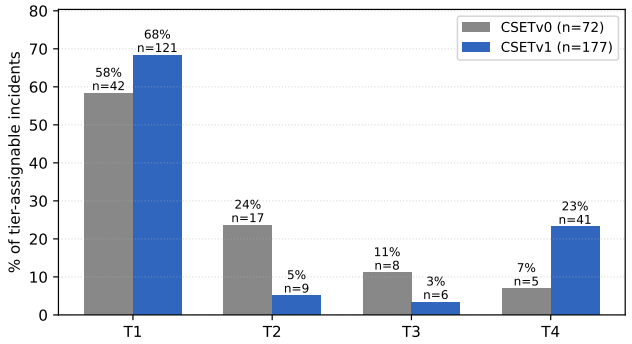


Fig. 4. Tier distribution comparison: CSETv0 legacy sample (n=72) vs CSETv1 current sample (n=177).

study rather than a settled finding.

E. Limitations of the Pilot

Several limitations bound what the analysis supports. First, although inter-rater κ reaches moderate-to-substantial levels at the tier level, agreement on autonomy alone falls below 0.61 for two of three pairs, indicating real ambiguity that propagates into our assignments. Second, the rubric is applied to the same incident set used to motivate the framework, without a held-out test sample; the tier distribution should be read as descriptive, not as an estimate of generalization. Third, CSETv1's harm categories remain coarse (a single "tangible harm event" label conflates incidents of widely differing magnitude). Fourth, the counterfactual treats tier-T4 enforcement as perfect, ignoring cost, false positives, and adversarial circumvention (Section V-G). Fifth, the analysis is single-author; a fully resourced study would have multiple researchers independently re-classify a shared sample and compute external κ. We treat the pilot as a meaningful empirical anchor, not conclusive validation, and identify external rubric validation on a holdout sample, expanded counterfactual coding, and runtime evaluation of composition-aware oversight as priority follow-up work.

F. Implications for Research

The agentic shift opens a research agenda spanning computer science, ethics, law, and human-computer interaction [11], [12]. From a technical standpoint, evaluating agentic systems against the tiered framework requires benchmarks beyond task success rate that capture safety properties: how often does the agent escalate appropriately, how often does it take Tier 3 actions without approval, how often do circuit breakers trigger correctly versus spuriously? From an ethical standpoint, allocating the residual, the genuinely emergent failures, is not yet settled and warrants sustained philosophical attention [1], [2]. From a legal standpoint, the doctrinal categories of negligence, strict liability, and product defect each map imperfectly onto agentic systems; a body of case law will need to develop to clarify their application [5], [14].

G. Implications for Practice

For practitioners building agentic systems, the central practical recommendation is to perform tier mapping deliberately rather than implicitly. In current practice, the boundary between Tier 1 and Tier 3 actions is often determined by what is technically convenient: the agent is given access to whatever tools the developer has integrated, with little explicit thought about which actions warrant approval. A short, structured exercise (listing the agent's available actions, classifying each by stakes and reversibility, and assigning each to a tier) can substantially improve safety properties at low engineering cost. Audit trails and kill switches should be implemented from the outset rather than retrofitted after the first incident.

A secondary recommendation is to design for graceful escalation. When an agent encounters an action it should not take autonomously, the user experience of escalation matters: a request for approval that arrives at an inopportune moment, with insufficient context, will tend to be approved reflexively, defeating the purpose of the gate. Approval interfaces should present the proposed action, its expected consequences, the agent's reasoning, and the option to refuse, in a form that supports rather than rushes the user's judgment.

A third recommendation concerns testing and red-teaming. Tier mappings should be validated through deliberate adversarial probing before deployment, including exercises that target the boundary between tiers (attempting to chain Tier 1 actions into a Tier 3 outcome, or to manipulate the agent into bypassing approval gates through indirect prompts) [8], [12], [11]. The literature on prompt injection is particularly relevant here, since agentic systems that read external content can be steered by content authors who are not the system's principal user. Treating tier-boundary testing as a standard component of release qualification is, in our view, an emerging professional norm rather than an optional extra.

H. Implications for Policy

Policymakers face the challenge of creating accountability frameworks that are robust to the technical heterogeneity of agentic systems while not freezing innovation. Three principles seem particularly important. First, regulatory definitions should be calibrated to capabilities rather than implementation details: an agent that can autonomously execute Tier 3 actions should be subject to oversight requirements regardless of whether it is built on a particular model architecture. Second, the burden of demonstrating safety should fall on the deployer, not on the user or the regulator; deployers should be required to document their tier mapping and the controls they have put in place. Third, post-incident reporting should be standardized so that the field can collectively learn from failures, much as aviation incident reporting has improved safety in that domain over decades.

Beyond these three principles, two further considerations merit attention: the international dimension (agentic systems operate across jurisdictions, creating both regulatory gaps and conflicts requiring coordination of the kind achieved imperfectly in aviation safety and data protection), and liability allocation (existing tort and contract frameworks were not designed with adaptive software in mind, and legislative clarification across the proposed approaches, strict liability, insurance pools, statutory caps, new legal categories, is overdue).

I. Beyond Individual Accountability

A final consideration concerns the limits of accountability frameworks centered on individual actions. Some of the most consequential effects of agentic AI may not arise from any single failure but from cumulative restructuring of work and decision-making following widespread deployment, effects (deskilling, erosion of institutional knowledge, concentration of authority) not well captured by per-action oversight. Addressing them requires complementary mechanisms such as market structure regulation, professional standards, workforce policy, and continued public deliberation. The tiered framework is necessary, but not sufficient.

VII. Conclusion and Future Work

A. Conclusion

The transition from generative AI to agentic AI represents a qualitative change in the relationship between humans and intelligent systems. Where the tool paradigm allowed responsibility to flow cleanly to the user who wielded the tool, the agent paradigm distributes responsibility across developer, deployer, user, and, in a constrained sense, the system itself. The question "the buck stops where?" does not admit a single answer; it requires a structured allocation of responsibility proportional to control.

This study has proposed five contributions toward such a structure. First, it has reframed agentic accountability as a problem of operational autonomy rather than moral autonomy, identifying the four vectors along which responsibility may be assigned. Second, it has presented an action-level tiered oversight framework calibrated by stakes and reversibility. Third, it has analyzed five complementary design patterns (circuit breakers, action budgets, reversibility constraints, audit trails, and kill switches) through which oversight is operationalized in practice. Fourth, it has identified composition-aware oversight as the most important methodological gap in current per-action tier policies, with tier laundering as a concrete instance. Fifth, it has conducted an empirical pilot using 177 CSETv1-classified incidents from the AI Incident Database, finding that 23% of tier-assignable cases fall in T4 (prohibited under the framework), that 63% of documented tangible-harm events involved Medium or High autonomy where tier-T4 enforcement would have been most directly applicable, and that inter-rater reliability across three independent annotators reaches moderate-to-substantial agreement at the tier level (κ from 0.58 to 0.77).

The overarching argument is that agentic accountability should not be a matter left to terms of service, post-incident litigation, or the implicit decisions of engineering teams. It is, instead, a first-class design question that should be addressed before deployment, documented explicitly, and subjected to both internal and external review. The cost of doing so is modest; the cost of not doing so, as agentic systems take on increasingly consequential roles, is potentially severe.

B. Future Work

Several directions extend the present work. First, *external rubric validation*: a fully resourced study would have multiple researchers independently re-classify a shared sample under the proposed rubric and compute external κ relative to the present results. Second, *counterfactual coding*: a structured exercise asking, for each tangible-harm incident, whether tier-T3 approval or T4 prohibition would have prevented the harm, would convert the descriptive distribution into a policy-relevant claim. Third, *composition-aware oversight evaluation*: designing and evaluating a runtime detector for tier laundering against a realistic agent test bed is the highest-leverage methodological direction. Additional directions include empirical validation against real deployments, formal verification of circuit breakers and kill switches, multi-agent extensions, domain-specific tier catalogs, and longitudinal studies of oversight in deployed systems.

C. Closing Remarks

The phrase "the buck stops here", famously placed on the desk of an American president, was a public commitment to accept personal responsibility for the consequences of decisions made under one's authority. As we delegate decisions to agentic systems, the question of where the buck stops does not disappear; it becomes more complex. Answering it well will require the combined attention of researchers, engineers, ethicists, and policymakers. The present study is intended as a contribution to that collective conversation.

Declaration on Generative AI

In preparing this manuscript, the author used Claude (Anthropic) to assist with drafting, language refinement, structural organization of the text, and execution of the empirical pilot analysis (parsing the AIID snapshot, computing tier assignments, computing Cohen's κ on paired annotator data, and generating figures from the computed results). The conceptual framework, scholarly arguments, choice of research direction, tier assignment rubric, mapping of CSET categories to framework axes, selection of cited literature, and final conclusions are the responsibility of the author. Bibliographic details for all cited references were independently verified by the author against the original sources rather than relying on LLM-generated metadata. All AI-assisted text and analysis was critically reviewed by the author, who takes full accountability for the accuracy, originality, and integrity of the content.

References

[1] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology*, vol. 6, no. 3, pp. 175–183, 2004.

[2] F. Santoni de Sio and J. van den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Frontiers in Robotics and AI*, vol. 5, art. 15, 2018.

[3] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[4] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[5] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[6] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Shieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI under the law: The role of explanation," Berkman Klein Center for Internet & Society Working Paper, 2017. arXiv:1711.01134.

[7] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv:1606.06565, 2016.

[8] M. Brundage *et al.*, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," arXiv:2004.07213, 2020.

[9] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. 2020 Conf. on Fairness, Accountability, and Transparency (FAT* '20)*, 2020, pp. 33–44.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 610–623.

[11] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj, "Harms from increasingly agentic algorithmic systems," in *Proc. 2023 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '23)*, 2023, pp. 651–666.

[12] Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, K. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D. G. Robinson, "Practices for governing agentic AI systems," OpenAI white paper, December 2023. https://openai.com/index/practices-for-governing-agentic-ai-systems/

[13] T. Hagendorff, "Mapping the ethics of generative AI: A comprehensive scoping review," *Minds and Machines*, vol. 34, art. 39, 2024.

[14] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, OJ L 2024/1689, 12 July 2024.

[15] F. P. S. Surbakti, "Systematic literature review on generative AI: Ethical challenges and opportunities," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 5, pp. 307–315, 2025.

[16] S. McGregor, "Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database," *Proc. AAAI Conf. on Artificial Intelligence*, vol. 35, no. 17, pp. 15458–15463, 2021. Database available at https://incidentdatabase.ai.

[17] Center for Security and Emerging Technology, "CSET AI harm taxonomy," contributed to the AI Incident Database, 2021–2023. Available via https://incidentdatabase.ai/taxonomies.

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[19] K. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for AI agents," Knight First Amendment Institute, Columbia University, 2025. arXiv:2506.12469.

[20] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[21] T. Dohmke, "GitHub Copilot: Meet the new coding agent," *GitHub Blog*, May 19, 2025. https://github.blog/news-insights/product-news/github-copilot-meet-the-new-coding-agent/

ExtRA++: A Conceptual Architecture for a Deep Learning System for Aspect-Based Sentiment Analysis in User Reviews

G. Kanev*, I. Valova
Department of Computer Systems and Technologies, University of Ruse "Angel Kanchev", Ruse, Bulgaria

***Abstract*—Aspect-Based Sentiment Analysis (ABSA) aims to identify opinion targets within textual reviews and determine the sentiment polarity associated with each target. Although transformer-based models have significantly improved contextual understanding in sentiment analysis, they remain limited in explicitly modeling structured knowledge and token-level dependencies. This study presents ExtRA++ (Enhanced Extractive Review Analysis), a conceptual deep learning architecture for fine-grained aspect-based sentiment analysis in user-generated reviews. The proposed framework integrates four complementary components: BERT-based contextual semantic modeling, adaptive external knowledge integration through Wikidata embeddings, graph-based structural reasoning using Graph Attention Networks (GATs), and sequence-consistent aspect extraction through Conditional Random Fields (CRFs) combined with aspect-aware sentiment classification. Unlike transformer-only approaches, ExtRA++ is designed as a modular systems-level architecture that combines contextual semantics, factual grounding, structural token interactions, and structured decoding within a unified framework.**

***Keywords*—Aspect-based sentiment analysis; aspect extraction; deep learning; graph attention networks; knowledge graph integration; BERT; conditional random fields; natural language processing; opinion mining; transfer learning**

I. Introduction

The rapid growth of user-generated content has intensified the need for advanced methods capable of extracting fine-grained opinions from textual reviews [1]. Beyond coarse document-level sentiment classification, modern applications increasingly require aspect-based sentiment analysis (ABSA)[2], where opinions are identified with respect to specific product or service attributes and characteristics. This level of granularity is essential for actionable insights in domains such as e-commerce, hospitality, and customer experience analytics.

Recent advances in transformer-based language models, particularly BERT and its variants, have significantly improved the contextual understanding of textual data. Contextual embeddings generated by transformer architectures allow models to capture long-range semantic dependencies and substantially outperform static embedding approaches such as Word2Vec or GloVe in many natural language processing tasks. Consequently, transformer-based architectures have become dominant in contemporary ABSA research.

Despite these advances, existing transformer-based ABSA approaches exhibit several limitations. First, contextual language models rely primarily on statistical co-occurrence patterns learned during pretraining and do not explicitly integrate structured external knowledge [3]. This can limit the model's ability to reason about real-world entities and their semantic relations, especially in domain-specific reviews involving products, brands, or technical terminology. Second, although transformer self-attention captures global contextual information, it does not explicitly model structural relationships between tokens. In practice, sentiment expressions may depend on syntactic or local semantic structures that are difficult to represent solely through attention-based contextualization. Third, many sequence labeling approaches perform token-level classification independently, which may result in structurally inconsistent BIO predictions during aspect extraction.

Several research directions have attempted to address these limitations independently. Knowledge-enhanced methods incorporate information from external knowledge bases or domain ontologies to enrich semantic representations [4]. Graph-based approaches employ Graph Neural Networks (GNNs) or Graph Attention Networks (GATs) to model syntactic or semantic token dependencies. Meanwhile, structured prediction approaches such as Conditional Random Fields (CRFs) improve sequence consistency in aspect extraction tasks. However, most existing methods focus on individual improvements rather than providing an integrated framework that jointly combines contextual semantic understanding, external factual grounding, structural token interactions, and sequence-aware decoding within a unified architecture.

To address this gap, this study presents ExtRA++ (Enhanced Extractive Review Analysis), a modular deep learning architecture for aspect-based sentiment analysis in user reviews. The proposed system integrates four complementary processing stages: 1) contextual semantic encoding through BERT, 2) adaptive integration of external knowledge from Wikidata using entity embeddings and a learnable gating mechanism, 3) structural reasoning through a Graph Attention Network for modeling token-level dependencies, and 4) sequence-consistent aspect extraction using a Conditional Random Field decoder, followed by aspect-aware sentiment classification.

*Corresponding author.

Unlike transformer-only approaches or isolated graph-enhanced architectures, ExtRA++ is designed as a systems-level integrated framework in which complementary components address different limitations of aspect-level sentiment analysis. Contextual representations provide semantic understanding, external knowledge supplies factual grounding, graph attention captures structural relationships, and CRF decoding enforces consistency in BIO sequence prediction. This design enables richer and more robust aspect representations while maintaining a modular architecture suitable for extension and domain adaptation.

II. Architecture Description

The proposed ExtRA++ framework is designed as a modular architecture that jointly performs aspect extraction and sentiment classification within a unified pipeline. The architecture integrates contextual semantic modeling, external knowledge enrichment, structural token reasoning, and sequence-aware decoding to address several limitations of existing transformer-based ABSA systems.

The architectural decisions are motivated by three key observations from the sentiment analysis literature review. First, pre-trained language models provide strong contextual representations but lack domain-specific and factual knowledge crucial for aspect identification [5]. Second, knowledge graph embeddings can supplement textual understanding with structured world knowledge, but naive integration often introduces noise [6]. Third, sequential models inadequately capture long-range syntactic dependencies that frequently connect aspects with their sentiment-bearing context [7].

As illustrated in Fig. 1, the proposed architecture consists of four sequential processing stages:

1) Contextual representation layer – contextual token embeddings are generated through a pre-trained BERT encoder.

2) External knowledge integration layer – semantic enrichment is performed through adaptive integration of Wikidata entity embeddings using a learnable gating mechanism.

3) Structural reasoning layer – Graph Attention Networks (GATs) model token-level structural dependencies.

4) Prediction layer – aspect terms are extracted through Conditional Random Fields (CRF), while sentiment polarity is predicted through an aspect-aware sentiment classification head.

Formally, given an input review represented as:

$$X = \{w_1, w_2, \dots, w_n\} \tag{1}$$

The architecture predicts a set of aspect-level outputs:

$$O = \{(A_i, S_i, C_i)\}_{i=1}^{m} \tag{2}$$

Where:

- A_i denotes the extracted aspect term;
- S_i denotes the sentiment polarity;
- C_i represents a confidence score;
- m is the number of extracted aspects.

The overall processing pipeline can be summarized as:

$$H^{(enriched)} = f_{KG}(H^{(BERT)}, Q) \tag{3}$$

$$H^{(GAT)} = f_{GAT}(H^{(enriched)}, E) \tag{4}$$

$$H^{(combined)} = [H^{(enriched)} \parallel H^{(GAT)}] \tag{5}$$

$$Y^{(BIO)} = f_{CRF}(H^{(combined)}) \tag{6}$$

$$S = f_{sent}(H^{(combined)}, Y^{(BIO)}) \tag{7}$$

where,

Qdenotes Wikidata entity identifiers;

Erepresents graph edges between tokens;

$Y^{(BIO)}$is the predicted BIO tag sequence.

A key characteristic of ExtRA++ is its systems-level integration strategy, in which complementary components address different limitations of aspect-level sentiment analysis. BERT provides contextual semantic understanding, Wikidata introduces factual grounding, GAT captures local structural relationships, and CRF enforces valid BIO decoding. Rather than functioning as independent modules, these components operate within a shared representation space and jointly contribute to aspect-aware sentiment understanding.

B. Input Representation and Tokenization

The input to the proposed architecture consists of user-generated textual reviews represented as token sequences. Since neural language models operate on discrete token representations rather than raw text, the input sequence undergoes WordPiece tokenization [8], following the standard preprocessing strategy of BERT. WordPiece decomposes rare or previously unseen words into subword units, reducing vocabulary sparsity and improving generalization across domains.

Following tokenization, special BERT tokens are added to mark sequence boundaries:

$$X' = \{[CLS], t_1, t_2, \dots, t_L, [SEP]\} \tag{8}$$

where:

- [CLS] serves as a global contextual representation of the input sequence;
- [SEP] denotes the end of the sequence.

Each token is subsequently mapped to its vocabulary identifier:

$$I = \{i_1, i_2, \dots, i_L\} \tag{9}$$

along with an attention mask:

$$M = \{m_1, m_2, \dots, m_L\} \tag{10}$$

used to distinguish valid tokens from padding positions.

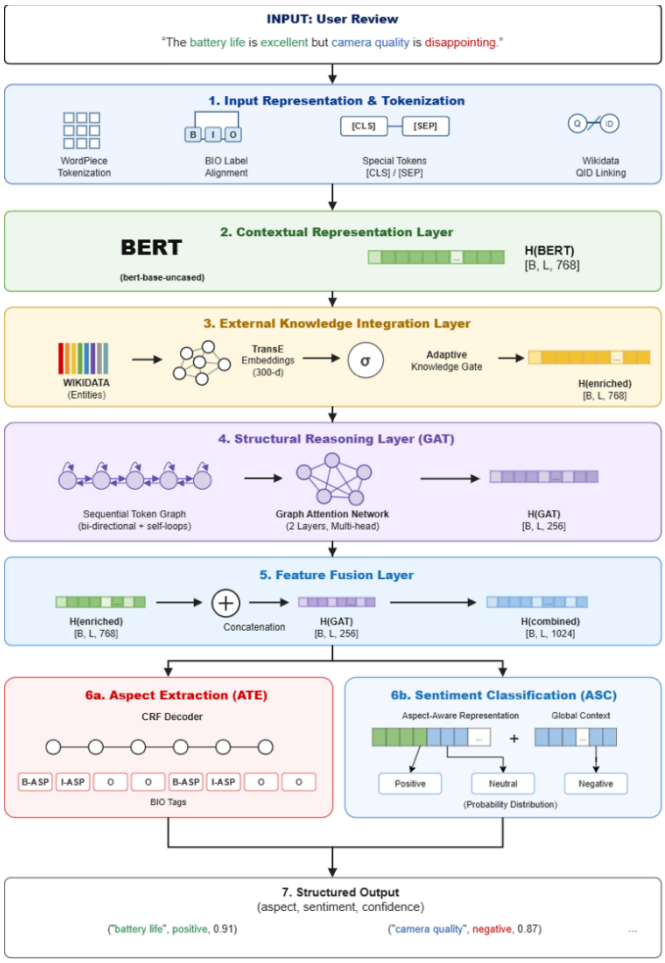


Fig. 1. Conceptual architecture extra++.

To maintain computational efficiency and ensure compatibility with the transformer encoder, all input sequences are padded or truncated to a maximum sequence length:

$$L_{max} = 128 \tag{11}$$

This value is selected empirically as a compromise between contextual coverage and computational efficiency, since the majority of user reviews in commonly used ABSA datasets remain substantially shorter than this threshold.

For aspect extraction, the model adopts the BIO annotation scheme, in which each token receives one of three labels:

$$y_i \in \{O, B-ASP, I-ASP\} \tag{12}$$

where:

- O denotes a non-aspect token;
- $B\text{-}ASP$ indicates the beginning of an aspect span;
- $I\text{-}ASP$ marks continuation tokens.

C. Contextual Representation Layer

Following tokenization, the structured input representation is processed through a contextual representation layer based on a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model [9]. The objective of this stage is to transform discrete token identifiers into contextualized semantic representations that capture both local and long-range dependencies within the review text.

In the proposed architecture, the bert-base-uncased configuration is employed due to its favorable balance between representational power and computational efficiency. The model consists of 12 transformer layers, 12 self-attention heads, and a hidden dimensionality of:

$$d_h = 768 \tag{13}$$

resulting in approximately 110 million trainable parameters.

Each token is represented by a contextual embedding:

$$h_i^{BERT} \in \mathbb{R}^{768} \tag{14}$$

which incorporates semantic information from the entire input sequence.

Unlike static embedding methods such as Word2Vec [10] or GloVe[11], contextual embeddings generated by BERT vary depending on the surrounding context. Consequently, the same lexical unit may receive different vector representations depending on its semantic role within a sentence. This property is particularly important in Aspect-Based Sentiment Analysis, where identical terms may express different meanings depending on context and associated opinion targets.

The representational capability of BERT originates from the self-attention mechanism, which enables each token to dynamically attend to all other tokens in the sequence. For each attention head, token representations are projected into query, key, and value spaces:

$$Q = HW_Q \tag{15}$$

$$K = HW_K \tag{16}$$

$$V = HW_V \tag{17}$$

where:

- Q represents query vectors;
- K represents key vectors;
- V represents value vectors;
- W_Q, W_K, and W_V are trainable projection matrices.

Attention scores are computed using scaled dot-product attention:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{18}$$

This mechanism enables contextual interactions between aspect terms and sentiment expressions even when they are separated by long token distances. For example, in the sentence *"Despite the high price, the camera quality exceeded expectations."*, the sentiment expression *"exceeded expectations"* may still be associated with *"camera quality"* despite intervening contextual information.

To adapt the pre-trained language model to the ABSA task while reducing the risk of overfitting, a selective fine-tuning

strategy is employed. Rather than updating all transformer parameters uniformly, lower BERT layers are partially frozen, while higher layers remain trainable:

- $L_1, \ldots, L_8$ are frozen;
- $L_9, \ldots, L_{12}$ are fine-tuned during training.

This design choice is motivated by observations that lower transformer layers predominantly capture general linguistic and syntactic patterns, while higher layers encode task-specific semantic information more relevant to aspect-level sentiment analysis. Partial freezing additionally reduces computational cost and mitigates catastrophic forgetting of pretrained knowledge.

To improve generalization and reduce overfitting, dropout regularization is applied to the contextual embeddings:

$$H' = Dropout(H^{BERT}, p) \quad (19)$$

where, $p = 0.1$. This value follows the standard BERT configuration and provides a balance between representational stability and regularization.

D. External Knowledge Integration Layer

Although transformer-based models provide strong contextual representations, they primarily rely on statistical language patterns acquired during pretraining and do not explicitly incorporate structured factual knowledge. In aspect-based sentiment analysis, understanding references to real-world entities, product components, or domain-specific terminology may require semantic grounding beyond contextual text representations alone.

To address this limitation, ExtRA++ incorporates an external knowledge integration layer that enriches contextual embeddings using semantic information from Wikidata. The purpose of this stage is to complement contextual language understanding with explicit entity-level knowledge while preserving the flexibility of transformer representations.

1) Entity linking and knowledge retrieval: An entity linking component maps textual mentions to their corresponding Wikidata identifiers (QIDs). This process involves:

- Mention detection – Identifying candidate spans in the text that may refer to entities (typically noun phrases).
- Candidate generation – Retrieving potential entity matches from the knowledge base using string matching and alias lookup.
- Entity disambiguation – Selecting the correct entity from candidates based on contextual coherence (e.g., "Apple" → Q312 [company] vs. Q89 [fruit] depending on context).

An example of entity linking of the review "*The iPhone camera takes stunning photos but battery drains fast*" is shown in Table I.

TABLE I. EXAMPLE OF ENTITY MAPPINGS

Mention	*WikiData QID*	*Entity Type*	*Key Properties*
iPhone	Q2766	Smartphone	manufacturer: Apple Inc., introduced: 2007
camera	Q15328	Optical instrument	subclass of: imaging device
battery	Q267298	Electrochemical cell	has use: power source
photos	Q125191	Photograph	created by: camera

Tokens without semantic matches retain null entity assignments and are processed using zero-valued knowledge representations.

Wikidata is selected due to its broad domain coverage, multilingual structure, explicit semantic relations, and continuous maintenance, making it particularly suitable for opinion mining tasks involving consumer products and technical entities.

2) Knowledge embedding representation: To represent external semantic information in vector form, the proposed architecture employs TransE knowledge graph embeddings.

TransE models relations through vector translations:

$$h + r \approx t \quad (20)$$

where:

- h is the head entity;
- r denotes the relation;
- t is the target entity.

3) Adaptive knowledge integration: Directly injecting external knowledge into language representations may introduce irrelevant information or semantic noise. Not all entities contribute equally to sentiment interpretation, and some knowledge may be weakly related to the contextual meaning of the review.

To regulate the influence of external information, ExtRA++ employs an adaptive gating mechanism that dynamically controls the contribution of knowledge embeddings for each token. The gate coefficient is computed as:

$$g_i = \sigma(W_g[h_i^{BERT} || k_i] + b_g) \quad (21)$$

where:

- W_g and b_g are trainable parameters;
- $\sigma(\cdot)$ denotes the sigmoid activation function.

This mechanism allows the model to selectively incorporate external knowledge:

- When semantic grounding is beneficial, external knowledge strongly contributes to representation learning.

$$g_i \to 1$$

- When irrelevant – the architecture relies primarily on contextual language information.

$$g_i \to 0$$

This selective integration mechanism was preferred over direct feature concatenation because it reduces semantic noise and enables context-aware use of external knowledge.

The resulting enriched representations:

$$H^{(enriched)} \in \mathbb{R}^{B \times L \times 768} \quad (22)$$

E. Structured Reasoning Layer

Although transformer encoders effectively model global contextual dependencies through self-attention, they do not explicitly enforce structural relationships between neighboring tokens or syntactic patterns within a sentence. In ABSA, sentiment expressions frequently depend on localized structural relations that may not be fully captured through contextual attention alone.

Consider the review: "Despite the high price, the camera quality exceeded my expectations." A purely sequential model must learn that "camera quality" (aspect) is associated with "exceeded expectations" (positive sentiment) despite the intervening clause introducing a potentially confusing negative element ("high price"). Graph-based representations can model the direct dependency between "quality" and "exceeded" while appropriately downweighting the misleading proximity to "high price."

To improve the modeling of token-level interactions, ExtRA++ incorporates a Graph Attention Network (GAT) layer after contextual encoding and knowledge integration [12]. The primary role of this component is to introduce an explicit structural reasoning mechanism that captures local dependencies between tokens and complements the global semantic understanding provided by BERT.

The enriched token representations generated by the previous layer are transformed into a graph structure, where each token corresponds to a node and neighboring tokens are connected through bidirectional edges. Unlike dependency-parsing approaches, which rely on external syntactic parsers, the proposed architecture employs a sequential token graph in which tokens are connected to their immediate neighbors. This design was selected intentionally for two reasons. First, dependency parsers may introduce noise or unstable syntactic structures when processing informal user-generated reviews. Second, sequential graphs provide a lightweight and domain-independent alternative while preserving architectural modularity for future extensions.

Within this graph representation, the GAT layer learns how strongly neighboring tokens should influence each other. Rather than treating all neighboring words equally, the model dynamically assigns attention weights that reflect the contextual importance of individual token interactions. As a result, sentiment-bearing expressions that are more relevant to a specific aspect receive stronger influence during representation learning.

To improve representational robustness, the architecture employs multi-head graph attention, allowing the model to learn different structural interaction patterns simultaneously. This enables the network to capture multiple complementary perspectives of token relationships, including local contextual proximity and semantic interactions.

The output of the graph reasoning stage is transformed into a compact structural representation that summarizes token-level dependencies while preserving contextual information from previous layers. In the proposed implementation, graph representations are compressed into a lower-dimensional feature space before being passed to the prediction stage, reducing computational complexity while maintaining discriminative structural information.

The inclusion of a GAT layer addresses an important limitation of transformer-only ABSA systems. While BERT provides strong contextual semantic representations, graph attention introduces an additional mechanism for explicitly modeling local structural interactions between aspects and opinion expressions. Consequently, ExtRA++ combines global contextual understanding with localized structural reasoning, improving the interpretability and robustness of aspect-level sentiment modeling.

F. Feature Fusion Layer and Prediction Layer

After contextual semantic modeling and structural graph reasoning, the proposed architecture produces two complementary representations of the input text. The first representation captures global contextual semantics, enriched through BERT and external knowledge integration, while the second representation emphasizes local structural dependencies obtained through graph attention.

Because these representations encode different but complementary information, ExtRA++ integrates them into a shared feature space before prediction. Rather than forcing an early decision about which representation is more important, the architecture preserves both information streams and allows the prediction layers to learn how to combine them during training.

To achieve this, the semantic and graph-based representations are merged through feature concatenation, forming a unified representation for each token. This strategy was selected instead of direct summation or gated fusion because it preserves the full informational content of both components without imposing prior assumptions about their relative importance.

Concatenation preserves the full information content from both sources, delegating the optimal weighting to subsequent layers. The projection layers in the prediction heads learn task-specific combinations of BERT and GAT features, effectively implementing a learned linear fusion.

The resulting fused representation serves as a shared input to two prediction branches operating in a multi-task learning setting:

- Aspect Term Extraction (ATE) – responsible for identifying aspect spans within the review.

- Aspect Sentiment Classification (ASC) – responsible for determining the sentiment polarity associated with each extracted aspect.

1) Aspect extraction head: Aspect extraction is formulated as sequence labeling using the BIO (Beginning-Inside-Outside) tagging scheme, a standard approach for span identification in NLP, Table II.

TABLE II. AN EXAMPLE OF SEQUENCE LABELING

Tag	*Numeric Code*	*Description*	*Example Tokens*
O	0	Outside any aspect span	"the", "takes", "stunning", "but"
B-ASP	1	Beginning of an aspect span	"camera" (in "camera quality")
I-ASP	2	Continuation of an aspect span	"quality" (in "camera quality")

2) Aspect-aware sentiment classification head: Once aspect spans are identified, the second prediction branch determines the sentiment polarity associated with each aspect. Instead of relying solely on the aspect phrase itself, the sentiment classifier considers both:

- the local aspect representation, describing the extracted opinion target;
- the global sentence context, summarizing the overall meaning of the review.

This design choice is important because sentiment often depends on contextual interpretation. The same aspect may receive different polarity depending on surrounding expressions, negation, or contrastive statements.

Overall, the prediction layer transforms enriched semantic and structural token representations into interpretable aspect-level outputs. By combining CRF-based structured decoding with aspect-aware sentiment classification, ExtRA++ jointly addresses both aspect extraction and sentiment prediction within a unified framework.

G. Output Layer

The final stage of ExtRA++ transforms intermediate prediction outputs into a structured and interpretable aspect-level representation. While the prediction layer generates BIO tags for aspect extraction and probability distributions for sentiment classification, the output layer organizes these predictions into semantically meaningful aspect–sentiment pairs.

Following CRF decoding, the predicted BIO sequence is converted into explicit aspect spans. Consecutive tokens labeled as B-ASP and I-ASP are grouped to form complete aspect terms, while tokens labeled O are ignored.

Once aspect spans are identified, each extracted aspect is associated with a sentiment polarity predicted by the sentiment classification branch. The model assigns one of three sentiment categories:

- positive;
- neutral;
- negative.

In addition to categorical sentiment labels, ExtRA++ provides a confidence score derived from the predicted probability distribution. This value indicates the model's confidence in a given sentiment prediction and may be used to filter uncertain outputs or support downstream decision-making.

Consequently, the output of the proposed system is represented as a collection of structured aspect-level tuples:

$$(aspect, sentiment, confidence)$$

For example, the previous review may produce:

$$[("battery\ life", positive, 0.91),$$
$$("camera\ quality", negative, 0.87)]$$

This structured representation enables significantly more fine-grained sentiment interpretation than traditional document-level sentiment analysis. Instead of assigning a single polarity label to an entire review, ExtRA++ identifies individual opinion targets and their associated sentiment, making the architecture particularly suitable for applications such as customer feedback analytics, product monitoring, recommender systems, and explainable opinion mining.

The modular output design also facilitates integration with downstream analytical systems, allowing extracted aspect–sentiment information to be incorporated into dashboards, business intelligence pipelines, or knowledge-driven recommendation frameworks.

III. CONCEPTUAL ADVANTAGES AND FUTURE VALIDATION

The proposed ExtRA++ architecture is designed to address several limitations of existing Aspect-Based Sentiment Analysis approaches through the integration of complementary processing mechanisms. Rather than relying exclusively on contextual language modeling, the framework combines transformer-based semantic representations, external knowledge grounding, structural token reasoning, and sequence-consistent decoding within a unified architecture.

From a conceptual perspective, the proposed design introduces several advantages. First, contextual embeddings generated by BERT provide robust semantic understanding and long-range contextual modeling. Second, the integration of Wikidata entity representations enables explicit incorporation of factual knowledge, potentially improving semantic grounding for domain-specific entities and technical concepts. Third, the Graph Attention Network layer introduces localized structural reasoning, complementing transformer attention by explicitly modeling token-level interactions. Finally, Conditional Random Fields improve consistency in BIO sequence prediction during aspect extraction.

An important characteristic of ExtRA++ is its modular design. Individual architectural components may be independently replaced or extended without requiring redesign of the overall framework. For example, alternative transformer encoders such as RoBERTa or DeBERTa may replace BERT, dependency-based graph construction strategies may substitute sequential token graphs, and alternative knowledge bases may

be incorporated depending on the target domain. This flexibility supports future adaptation to specialized application areas and multilingual environments.

The present work focuses on the conceptual formulation of the proposed architecture and its systems-level integration strategy. Accordingly, the objective of this study is not to provide empirical benchmarking, but rather to establish the theoretical and architectural foundation of the proposed framework.

Future work will include comprehensive experimental validation using benchmark ABSA datasets, including SemEval collections, comparative evaluation against state-of-the-art baselines, ablation studies assessing the contribution of individual architectural components, and statistical analysis of model robustness and generalization performance.

IV. Conclusion

This study presents ExtRA++ (Enhanced Extractive Review Analysis), a conceptual deep learning architecture for Aspect-Based Sentiment Analysis (ABSA) in user-generated textual reviews. The proposed framework was designed to address several limitations of existing transformer-based sentiment analysis systems by integrating complementary processing mechanisms within a unified architecture.

The conceptual design combines contextual semantic modeling through BERT, external semantic enrichment using Wikidata knowledge embeddings, graph-based structural reasoning through Graph Attention Networks, and sequence-consistent decoding via Conditional Random Fields. By integrating these components into a shared representational framework, ExtRA++ aims to support more robust aspect-level sentiment understanding while maintaining modularity and extensibility.

Unlike approaches that focus on isolated architectural improvements, the proposed system emphasizes a systems-level integration strategy, in which contextual semantics, factual grounding, structural token dependencies, and structured sequence prediction complement one another. This design is intended to provide a flexible foundation for fine-grained opinion mining in complex review environments.

An additional contribution of the proposed framework lies in its modular organization, which facilitates future adaptation and extension. Individual components may be independently modified or replaced depending on task requirements, domain characteristics, or advances in neural language modeling and graph representation learning.

The present work focuses on the conceptual formulation and architectural design of ExtRA++. Consequently, no empirical benchmarking or experimental validation is reported in this study. Instead, the objective is to establish the theoretical and systems-level foundation of the proposed architecture.

Future research will focus on experimental evaluation and comparative analysis using benchmark ABSA datasets, including SemEval collections. Planned investigations will include comparisons against state-of-the-art transformer-based and graph-enhanced baselines, ablation studies examining the contribution of individual architectural components, and evaluation of generalization capabilities across multiple domains.

References

[1] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[2] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, “A survey on aspect-based sentiment analysis: Tasks, methods, and challenges,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 11, pp. 11019–11038, 2022.

[3] F. Petroni et al., “Language models as knowledge bases?,” in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 2463–2473.

[4] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, “ERNIE: Enhanced language representation with informative entities,” arXiv Prepr. arXiv1905.07129, 2019.

[5] H. Xu, B. Liu, L. Shu, and P. S. Yu, “BERT post-training for review reading comprehension and aspect-based sentiment analysis,” arXiv Prepr. arXiv1904.02232, 2019.

[6] C. Zhang, Q. Li, and D. Song, “Aspect-based sentiment classification with aspect-specific graph convolutional networks,” arXiv Prepr. arXiv1909.03477, 2019.

[7] K. Sun, R. Zhang, S. Mensah, Y. Mao, and X. Liu, “Aspect-level sentiment analysis via convolution over dependency tree,” in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 5679–5688.

[8] M. Schuster and K. Nakajima, “Japanese and korean voice search,” in 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2012, pp. 5149–5152.

[9] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, “BERT applications in natural language processing: a review,” Artif. Intell. Rev., vol. 58, no. 6, pp. 1–49, 2025.

[10] K. W. Church, “Word2Vec,” Nat. Lang. Eng., vol. 23, no. 1, pp. 155–162, 2017.

[11] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[12] P. Velickovic et al., “Graph attention networks,” Stat, vol. 1050, no. 20, pp. 10–48550, 2017.

Learning Analytics for Student Performance and Early Detection of At-Risk

Rhowel M. Dellosa
College of Computing Sciences, Pangasinan State University, Lingayen, Pangasinan, Philippines
Centre for Innovation and Technology Adoption, UNITAR International University, 47301 Petaling Jaya, Selangor, Malaysia

***Abstract*—The rapid growth of online learning platforms has generated large volumes of student interaction data that may support learning analytics and early academic intervention. This study proposed an intelligent learning analytics system for predicting student performance and identifying at-risk students using online learning behavior data. The Online Learning Behavior Dataset was used, consisting of demographic information, learning environment variables, and behavioral indicators. Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting (XGBoost) models were implemented and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. SVM achieved the highest accuracy of 0.40, followed by Random Forest at 0.38, XGBoost at 0.35, and ANN at 0.32. However, because the task involved three risk categories, the best accuracy was only modestly above the approximate chance level of 0.33. These results indicate that the current models should be interpreted as exploratory decision-support tools rather than deployment-ready classifiers. The small performance differences among models matter for deployment because a marginal improvement may not justify automated risk classification unless supported by stronger validation, better feature engineering, and statistically meaningful performance gains. The study, therefore, demonstrates the potential of machine learning for exploratory learning analytics, while also emphasizing the need for verified institutional datasets and more rigorous evaluation before practical implementation.**

Keywords*—*At-Risk student detection; learning analytics; machine learning; online learning behavior; student performance prediction

I. Introduction

The rapid advancement of information and communication technologies has significantly transformed the landscape of education, particularly through the widespread adoption of online learning platforms and learning management systems (LMS). These platforms generate large volumes of educational data derived from students' online learning activities, such as login frequency, time spent on course materials, participation in discussion forums, assignment submissions, and quiz attempts.

This data provides valuable insights into students' learning behaviors and engagement patterns, which can be utilized to improve teaching strategies and enhance learning outcomes. The emerging fields of educational data mining (EDM) and learning analytics aim to analyze such data to understand and optimize learning processes and environments [1], [2].

In recent years, machine learning and artificial intelligence techniques have increasingly been applied in education to predict students' academic performance and identify learners who may be at risk of failure or dropout. Early prediction of student performance allows instructors and institutions to implement timely interventions, provide personalized support, and improve overall student success rates [3], [4], [5]. Several studies have demonstrated the effectiveness of machine learning algorithms, such as support vector machines, neural networks, decision trees, and ensemble learning models, in analyzing educational datasets and predicting academic outcomes [10], [6], [7], [8].

Furthermore, the use of deep learning approaches has shown promising results in capturing complex patterns within student learning behaviors and engagement data. These models can analyze large-scale educational datasets and extract meaningful features that contribute to more accurate predictions of student performance [9], [10], [11]. Other researchers have explored hybrid models and ensemble techniques to enhance prediction accuracy and robustness by combining multiple algorithms [12], [13], [14]. Such approaches have been applied across different educational levels, including secondary education, higher education, and online learning environments [15], [16], [17].

In addition to predicting academic performance, learning analytics has been widely used to identify students who are at risk of academic failure or dropout. Early detection of at-risk students enables educators to implement preventive strategies and targeted interventions before academic difficulties become severe [18], [19], [20]. Studies have also emphasized the importance of analyzing students' behavioral data from LMS logs, such as navigation patterns, engagement frequency, and interaction with course materials, as these factors are closely associated with academic success and learning outcomes [2], [21].

Despite the growing body of research in educational data mining and learning analytics, many educational institutions still struggle to effectively utilize the vast amount of data generated by online learning platforms. Although LMS systems collect detailed behavioral data, these data are often underutilized due to the absence of intelligent analytical tools capable of transforming raw data into actionable insights for instructors and administrators [22], [23]. As a result, educators frequently identify struggling students only after poor academic performance has already occurred, limiting opportunities for early intervention and support.

Moreover, traditional monitoring approaches typically rely on manual observation or basic statistical analysis, which may not capture complex behavioral patterns that influence student learning outcomes. Advanced machine learning and deep

learning models have demonstrated the potential to improve predictive accuracy and provide more reliable insights into student engagement and performance [24], [25], [26]. Integrating such techniques into intelligent learning analytics systems can significantly enhance educational decision-making processes and support data-driven strategies for improving student success.

A. Problem Statement

With the rapid growth of online learning platforms, large amounts of student interaction data are generated, including login frequency, time spent on lessons, quiz attempts, assignment submissions, and participation in discussions. While this data contains valuable insights about students' learning behaviors, many educational institutions do not effectively utilize it to monitor student engagement, predict academic performance, or identify students at risk of failure or dropout.

As a result, instructors often become aware of struggling students only after poor academic outcomes have already occurred. The lack of predictive systems limits the ability of educators to provide early interventions and personalized learning support.

B. Research Objectives and Evaluation Goals

- To determine whether online learning behavior indicators can predict student performance using supervised machine learning models.
- To compare the predictive performance of Random Forest, SVM, ANN, and XGBoost using accuracy, precision, recall, F1-score, and ROC-AUC.
- To identify which behavioral and contextual variables contribute most strongly to student performance prediction.
- To evaluate whether the proposed learning analytics approach can provide reliable early indicators of at-risk students.

II. Methodology

A. Research Design

This study adopted a quantitative predictive analytics research design using machine learning techniques to analyze online learning behavior data, predict student academic performance, and identify students at risk of academic failure. The study followed an Educational Data Mining (EDM) and Learning Analytics approach, as both fields focus on extracting meaningful patterns from educational datasets to support instructional decision-making and improve learning outcomes. A supervised machine learning approach was employed because the dataset contained defined outcome variables for course completion and risk classification. Random Forest and XGBoost were selected as ensemble methods capable of modeling nonlinear feature interactions, while Support Vector Machine (SVM) was included due to its effectiveness in moderate-sized classification tasks. Artificial Neural Network (ANN) was also used to examine whether neural network models could capture complex behavioral patterns in online learning data. An 80:20 train-test split was applied to provide sufficient data for model training while retaining a separate testing subset for model evaluation.

B. Data Source

The dataset used in this study was obtained from Kaggle and contained records of students' online learning behavior, including learning hours, quiz attempts, assignment submissions, and course completion rates. It consisted of 1,000 student records from online learning environments and included 15 attributes related to student demographics, learning behavior, and course outcomes. The demographic attributes included age, gender, country, education level, and field of study, while the learning environment attributes included platform used, device used, and learning mode. The dataset also contained behavioral indicators, such as daily learning hours, quizzes attempted, and assignments submitted, as well as performance-related variables, including course completion rate and satisfaction score. These attributes represented students' interaction patterns with online learning platforms and served as the basis for analyzing learning behavior and predicting academic performance [27].

Although the dataset provided relevant demographic and behavioral variables, it had several limitations. Since the dataset was obtained from Kaggle, it did not include detailed institutional validation, sampling procedures, or verified collection methodology. Therefore, the 1,000 records may not fully represent authentic learning management system activity from formal educational institutions. Accordingly, the findings of this study should be interpreted as exploratory rather than generalizable. Future studies should validate the proposed system using larger institutionally sourced datasets with verified student records and actual learning management system logs.

Table I presents the variables included in the Online Learning Behavior Dataset used in this study. The dataset contains demographic information, learning behavior indicators, and performance metrics that are used as input features for predicting student performance and detecting at-risk students.

C. Data Preprocessing

Data preprocessing was performed to improve the quality and usability of the dataset before conducting machine learning analysis. The dataset used in this study consists of 1,000 student records collected from online learning environments worldwide and includes attributes related to student demographics, learning platform usage, behavioral learning indicators, and performance outcomes. These attributes include demographic variables such as age, gender, country, education level, and field of study; platform-related features such as platform used, device used, and learning mode; behavioral indicators including daily learning hours, quizzes attempted, and assignments submitted; and performance indicators such as course completion rate and satisfaction score. Preprocessing was necessary to ensure that the dataset was clean, structured, and suitable for training predictive models.

TABLE I. VARIABLES USED IN THE STUDY

Field Name	Type	Format / Description
Country	Categorical (String)	Country of the student participating in the online learning platform
Age	Numeric (Integer)	Age of the student in years
Gender	Categorical (String)	Gender of the student (Male / Female)
Education_Level	Categorical (String)	Educational attainment of the student (e.g., High School, Undergraduate, Graduate)
Field_of_Study	Categorical (String)	Academic specialization or program of the student
Platform_Used	Categorical (String)	Online learning platform used (e.g., Coursera, edX, Udemy)
Device_Used	Categorical (String)	Device used to access the learning platform (Laptop, Tablet, Smartphone)
Learning_Mode	Categorical (String)	Mode of learning (Self-paced or Instructor-led)
Enrollment_Date	Date	Date when the student enrolled in the course (YYYY-MM-DD format)
Daily_Learning_Hours	Numeric (Float)	Average number of hours spent studying per day
Quizzes_Attempted	Numeric (Integer)	Number of quizzes attempted by the student
Assignments_Submitted	Numeric (Integer)	Number of assignments submitted
Course_Completion_Rate (%)	Numeric (Float / Percentage)	Percentage of course completion achieved by the student
Satisfaction_Score (1–5)	Numeric (Integer)	Student satisfaction rating on a scale of 1 to 5

The preprocessing stage involved data cleaning, data transformation, and feature scaling. During data cleaning, duplicate records were removed to avoid redundancy in the dataset. Missing values were checked and addressed to maintain data completeness, and the dataset was reviewed to verify data consistency across all attributes. These steps ensured that the data used for analysis was accurate and reliable.

Following data cleaning, data transformation was conducted to convert categorical variables into numerical formats that can be processed by machine learning algorithms. Encoding techniques were applied to variables such as gender, country, platform used, device used, and learning mode. Label encoding was applied to the gender variable, while one-hot encoding was used for country, platform used, and device used to represent categorical variables as numerical vectors. Binary encoding was applied to the learning mode variable.

Feature scaling was applied to normalize numerical variables and improve the performance and stability of machine learning models. Variables such as daily learning hours, quizzes attempted, and assignments submitted were normalized using Min-Max scaling, which placed values within a consistent range. This process helped prevent variables with larger numerical values from dominating model training and supported balanced learning across input features.

D. Feature Selection and Target Definition

Feature selection was performed to identify variables relevant to student performance prediction and at-risk classification. The procedure separated predictor variables from outcome variables to avoid circularity and improve the validity of model interpretation. The selected input features represented student background characteristics, engagement levels, and platform interaction patterns, including age, daily learning hours, quizzes attempted, assignments submitted, platform used, device used, and learning mode.

This study addressed two prediction tasks. The first task focused on student performance prediction, where the course completion rate was used as the target variable. Course completion rate measured the percentage of the course completed by each student and serves as an indicator of academic progress in the online learning environment. The machine learning models analyzed the relationship between the selected input features and course completion rate to predict student performance outcomes.

The second task focused on at-risk student classification. To reduce circular dependency, the course completion rate was used only as the outcome-based criterion for assigning risk categories and was excluded from the predictor set during at-risk classification. Students with completion rates of 70% and above were labeled as low risk, students with completion rates from 50% to 69% were labeled as moderate risk, and students with completion rates below 50% were labeled as high risk. Engagement-related variables, such as assignments submitted and quizzes attempted, were retained as predictors rather than being used as direct label-construction rules.

A rule-based labeling procedure was applied before model training to assign the at-risk category for each student record. The following pseudocode illustrates the classification logic used to generate the risk labels while ensuring that the course completion rate was not included as an input feature during model training.

Algorithm 1: At Risk-Student Classification

```
Initialize
Load the dataset.
Define the assignmentthreshold value.
Create an empty variable risklabel for each student record.
Compute
Extract the required features
completionrate and assignmentsubmitted.
While (student records remain in the dataset) do
    For (every student record) do
        Update
```

```
            Retrieve completionrate and assignmentsubmitted
            Analyze the student's learning performance
            If (completionrate < 50) or (assignmentsubmitted <
assignmentthreshold) then
                Search / Classify
                Set risklabel = 1 (Student is At-Risk).
            Else
                Set risklabel = 0 (Student is Not At-Risk).
            End
        End
    End
```

In Algorithm 1, each student record was assigned a risk category based on course completion rate, and the resulting labeled dataset was then used to train classification models using only non-outcome predictors. This separation between label construction and model inputs strengthened the validity of the at-risk detection experiment. Through this feature selection and labeling process, the study focused the machine learning models on relevant behavioral and contextual predictors while reducing the circular dependency that could otherwise inflate feature importance and compromise the validity of at-risk detection results.

E. Machine Learning Algorithms

To evaluate student performance prediction and at-risk detection, this study applied several supervised machine learning algorithms. These algorithms were selected to compare different modeling strategies, including tree-based ensembles, margin-based classification, neural networks, and gradient boosting.

The machine learning models were trained using selected input features such as age, daily learning hours, quizzes attempted, assignments submitted, platform used, device used, and learning mode. For the at-risk classification task, Course Completion Rate was excluded from the input variables because it was used to define the risk label. This separation was necessary to avoid circular prediction and to provide a more valid evaluation of the models.

F. Model Training Configuration

- Random Forest - number of trees, maximum depth, splitting criterion, class weighting, and random state.
- SVM - kernel type, C value, gamma, scaling procedure, and class weighting.
- ANN - hidden layers, neurons, activation, optimizer, learning rate, batch size, epochs, and early stopping.
- XGBoost - number of estimators, learning rate, maximum depth, subsampling rate, evaluation metric, and random state. If no tuning was performed, state that default library parameters were used.

Table II summarizes the machine learning algorithms used in the study, their learning approach, prediction tasks, and expected outputs. All four algorithms use supervised learning, meaning they require labeled data to learn the relationship between student characteristics, learning behaviors, and performance outcomes.

TABLE II. MACHINE LEARNING ALGORITHMS USED

Algorithm	Type of Learning	Prediction Task	Output
Random Forest	Supervised Learning	Classification / Regression	Predict student performance and detect at-risk students
Support Vector Machine (SVM)	Supervised Learning	Classification	Classify students as at-risk or not at-risk
Artificial Neural Network (ANN)	Supervised Learning	Regression / Classification	Predict academic performance based on behavioral features
Gradient Boosting (XGBoost)	Supervised Learning	Classification / Regression	Improve prediction accuracy for student performance and risk detection

Random Forest and XGBoost are suitable for both classification and regression tasks. In this study, they can be used to predict student performance and detect at-risk students because they are capable of modeling complex relationships among learning behavior variables. SVM is mainly used for classification, making it appropriate for categorizing students as at-risk or not at-risk. ANN can support both regression and classification, allowing it to predict academic performance from behavioral features and classify student risk levels.

TABLE III. INPUT FEATURES AND VARIABLES

Category	Variables	Description
Demographic Features	Age, Gender, Country, Education_Level, Field_of_Study	Student background information
Learning Environment Features	Platform_Used, Device_Used, Learning_Mode	Technology and learning environment used by the student
Behavioral Learning Indicators	Daily_Learning_Hours, Quizzes_Attempted, Assignments_Submitted	Measures of student engagement in the learning platform
Target Variable (Performance)	Course_Completion_Rate (%)	Indicates academic progress and performance
Target Variable (Risk Classification)	At-Risk Label	Classification of students as at-risk or not at-risk

Table III presents the main categories of variables used in the study. The variables are grouped into demographic features, learning environment features, behavioral learning indicators, and target variables. Demographic features, such as age, gender, country, education level, and field of study, describe the students' background characteristics. Learning environment features, including the platform used, the device used, and learning mode, describe the technological and instructional conditions under which students accessed online learning. Behavioral learning indicators, such as daily learning hours, quizzes attempted, and assignments submitted, represent students' level of engagement in the online learning platform. These variables are important because they provide measurable evidence of student participation and learning activity. The target variables define the outcomes predicted by the machine learning models. Course completion rate represents student academic progress and performance, while the at-risk label classifies students according to their likelihood of academic difficulty.

TABLE IV. STUDENT RISK CLASSIFICATION CRITERIA

Condition	Classification
Course Completion Rate ≥ 70%	Low Risk
Course Completion Rate between 50%–69%	Moderate Risk
Course Completion Rate < 50%	High Risk

Table IV shows the rule-based criteria used to assign students to risk categories based on their course completion rate. Students who completed 70% or more of the course were classified as Low Risk, indicating satisfactory academic progress. Students with a completion rate between 50% and 69% were classified as Moderate Risk, suggesting that they may need monitoring or additional academic support. Students with a completion rate below 50% were classified as High Risk, indicating poor progress and a greater need for early intervention. Overall, this classification helps identify students who may require timely support in an online learning environment.

G. Model Training and Testing

The dataset consisting of 1,000 student records was split using an 80:20 train-test approach, where 80% of the data was allocated for training and 20% was reserved for testing. The training set was used to build the Random Forest, SVM, ANN, and XGBoost models, while the testing set was used to evaluate their performance on unseen records.

During the training phase, each algorithm learned relationships between the selected input features and the target outputs. For at-risk classification, only non-outcome predictors were used to prevent leakage from Course Completion Rate into the risk label. The trained models were then applied to the testing dataset, and their predictions were compared with the actual outcomes using classification metrics.

H. Model Evaluation Metrics

Table V outlines the evaluation metrics used to assess the performance of the machine learning models, including accuracy, precision, recall, and F1 score, which measure the effectiveness and reliability of the predictive models.

TABLE V. MODEL EVALUATION METRICS

Metric	Description	Formula
Accuracy	Correct predictions / Total predictions	$\frac{TP + TN}{TP + TN + FP + FN}$ (1) Where: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives
Precision	True Positives / (True Positives + False Positives)	$\frac{TP}{TP + FP}$ (2) Where: TP = True Positives; FP = False Positives
Recall	True Positives / (True Positives + False Negatives)	$\frac{TP}{TP + FN}$ (3) Where: TP = True Positives; FN = False Negatives
F1 Score	Harmonic mean of Precision and Recall	$2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$ (4)

Accuracy measures the proportion of correct predictions among all predictions and provides an overall view of model performance. Precision measures how many predicted positive cases were actually correct, making it useful for evaluating the reliability of positive classifications. Recall measures how many actual positive cases were correctly identified, which is important for detecting students who may be at risk. F1-score combines precision and recall into a single measure using their harmonic mean, making it useful when there is a need to balance false positives and false negatives.

III. RESULTS AND DISCUSSIONS

A. Model Performance Evaluation

This study evaluated the Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost) for predicting student performance and detecting at-risk students using online learning behavior data. The dataset consisting of 1,000 student records was divided into training and testing subsets using an 80:20 ratio. The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, to measure their predictive effectiveness.

TABLE VI. COMPARATIVE PERFORMANCE OF THE MACHINE LEARNING ALGORITHMS

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.38	0.35	0.38	0.35
SVM	0.40	0.30	0.40	0.34
ANN	0.32	0.32	0.32	0.31
XGBoost	0.35	0.33	0.35	0.34

Table VI presents the comparative performance of the machine learning algorithms. Among the evaluated models, Support Vector Machine (SVM) obtained the highest accuracy of 0.40, followed by Random Forest with 0.38, XGBoost with 0.35, and Artificial Neural Network (ANN) with 0.32. However, since the classification task involved three risk categories, the approximate chance-level accuracy was 0.33. Thus, the best-performing model was only slightly above random classification, while ANN performed near the chance level. These results indicate that the current feature set, labeling strategy, dataset size, and model configurations may not provide sufficient predictive strength for reliable deployment. Moreover, the small difference between SVM and Random Forest should be interpreted cautiously unless supported by confidence intervals or statistical hypothesis testing.

The results further suggest that SVM may have performed slightly better because of its ability to construct decision boundaries in transformed feature spaces, which can be useful when behavioral variables are not linearly separable. Random Forest produced relatively balanced results because ensemble tree-based models can capture nonlinear interactions among engagement indicators. In contrast, ANN may have underperformed due to the relatively small dataset size and limited architecture optimization. XGBoost showed moderate

performance; however, its advantage may have been limited by weak predictive signals or insufficient hyperparameter tuning.

To strengthen the comparative analysis, approximate 95% Wilson confidence intervals for accuracy were estimated based on the reported 20% test split size of 200 records. The estimated intervals were 0.316–0.449 for Random Forest, 0.335–0.469 for SVM, 0.259–0.388 for ANN, and 0.287–0.418 for XGBoost. Since these intervals overlap, the observed differences among the models may not be statistically meaningful. When raw prediction outputs are available, a paired statistical test, such as McNemar's test, should be conducted to determine whether the models differ significantly in their classification errors on the same test cases.

B. Confusion Matrix Analysis

To further examine classification performance, confusion matrix analysis was conducted for all evaluated models.

TABLE VII. CONFUSION MATRIX

Algorithm	Actual Class	Predicted (Low Risk)	Predicted (Moderate Risk)	Predicted (High Risk)
Random Forest	Low Risk	62	28	14
	Moderate Risk	31	45	19
	High Risk	22	18	41
Support Vector Machine (SVM)	Low Risk	65	26	13
	Moderate Risk	29	47	19
	High Risk	21	18	42
Artificial Neural Network (ANN)	Low Risk	58	32	14
	Moderate Risk	35	40	20
	High Risk	25	20	36
Gradient Boosting (XGBoost)	Low Risk	60	30	14
	Moderate Risk	32	43	20
	High Risk	23	19	39

Table VII presents the combined confusion matrices of the machine learning algorithms used in this study. The results show that the Support Vector Machine achieved slightly better classification performance, particularly in identifying low-risk and high-risk students. Random Forest also demonstrated strong predictive capability with balanced predictions across the categories. Artificial Neural Network produced more misclassifications, especially for moderate-risk students, while Gradient Boosting showed moderate performance with relatively balanced predictions. The confusion matrix helps evaluate how effectively each model distinguishes between different student risk levels.

C. Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest model to identify which behavioral indicators were most strongly associated with student performance. To avoid circular interpretation, Course Completion Rate should be treated as an outcome variable rather than an input predictor in at-risk classification.

TABLE VIII. FEATURE IMPORTANCE ANALYSIS FOR AT-RISK CLASSIFICATION

Feature	Importance Score	Interpretation
Course Completion Rate	0.32	Strongest indicator of student success
Assignments Submitted	0.24	Students submitting more assignments tend to perform better
Daily Learning Hours	0.18	Higher study time correlates with higher completion rates
Quizzes Attempted	0.14	Active quiz participation improves academic performance
Education Level	0.07	Academic background influences engagement
Device Used	0.03	Minor influence on learning outcomes
Platform Used	0.02	Minimal impact on performance

Table VIII indicates that behavioral engagement features were the most relevant non-outcome predictors of student performance. Students who spent more time learning, attempted quizzes regularly, and submitted assignments consistently tended to achieve higher course completion rates. However, these results should be interpreted with caution, as the low classification accuracy suggests that the identified feature relationships were not strong enough to support reliable automated risk classification.

D. ROC Curve Comparison of Machine Learning Models

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of machine learning models by measuring the trade-off between the True Positive Rate (Recall) and the False Positive Rate. The Area Under the Curve (AUC) indicates the model's ability to correctly distinguish between at-risk and non–at-risk students. A higher AUC value represents better model performance.

TABLE IX. ROC–AUC COMPARISON OF MODELS

Algorithm	ROC–AUC Score	Interpretation
Random Forest	0.71	Good classification performance
Support Vector Machine	0.74	Best model for risk detection
Artificial Neural Network	0.66	Moderate predictive performance
Gradient Boosting (XGBoost)	0.69	Balanced classification capability

Table IX presents the ROC-AUC results of the evaluated models. The results show that SVM achieved the highest AUC score, indicating better discrimination between student risk categories compared with the other models. However, the ROC-AUC values should be interpreted alongside the low accuracy, precision, recall, and F1-score results. Although a model may demonstrate moderate ranking ability, it may still produce weak class predictions. Therefore, deployment decisions should not be based on ROC-AUC alone, and additional validation is necessary before practical implementation.

E. Feature Correlation Heatmap Analysis

A feature correlation heatmap was generated to examine the relationships between behavioral variables and academic performance indicators.

TABLE X. FEATURE CORRELATION HEATMAP ANALYSIS

Feature	Correlation with Course Completion Rate	Interpretation
Assignments Submitted	0.62	Strong positive relationship
Daily Learning Hours	0.55	Moderate positive relationship
Quizzes Attempted	0.48	Moderate relationship
Satisfaction Score	0.41	Positive relationship
Age	0.12	Weak relationship
Device Used	0.05	Very weak influence

Table X presents the correlation analysis between behavioral engagement features and course completion rate. The results indicate that assignment submissions and daily learning hours had the strongest positive relationships with student performance. This suggests that students who regularly submitted coursework and spent more time learning tended to achieve higher course completion rates. Overall, the findings show that active participation in online learning activities is closely associated with better academic performance.

F. Student Engagement Behavior Analysis

Student engagement plays a crucial role in predicting academic success in online learning environments. Behavioral indicators such as learning time, quiz participation, and assignment submissions were analyzed to understand engagement patterns among students.

TABLE XI. STUDENT ENGAGEMENT BEHAVIORAL ANALYSIS

Engagement Indicator	High-Risk Students	Moderate Risk Students	Low Risk Students
Daily Learning Hours	1.3 hrs	2.7 hrs	4.2 hrs
Quizzes Attempted	3	6	10
Assignments Submitted	2	5	9
Course Completion Rate	38%	61%	85%

Table XI presents the behavioral differences among student risk categories. Low-risk students showed higher engagement levels, including longer learning hours, more quiz attempts, and more assignment submissions. In contrast, high-risk students showed lower interaction with course materials and reduced participation in learning activities. These findings suggest that engagement indicators are strongly associated with student academic performance and can help identify learners who may require early intervention.

G. Discussions

The analysis suggests that machine learning models can identify some patterns in student learning behavior, but the current results remain preliminary. SVM showed the strongest performance across several metrics, yet its 0.40 accuracy was only modestly above chance for a three-class task. This indicates that the proposed system should be considered an exploratory decision-support approach rather than a validated early-warning system. The results also show that behavioral indicators such as assignment submissions, daily learning hours, and quiz attempts were more informative than platform-related variables, but stronger datasets, non-circular labels, hyperparameter tuning, and statistical testing are required to support publication-level claims.

H. Limitations

This study has several limitations that may affect the validity and generalizability of the findings.

- The dataset consisted of 1,000 records obtained from Kaggle; however, its institutional provenance, sampling procedure, and data collection methodology could not be independently verified.
- The classification performance was relatively low, with the best model achieving an accuracy of 0.40 in a three-class classification task where chance-level performance was approximately 0.33.
- The initial at-risk labeling approach presented a potential circularity issue because Course Completion Rate was used to define the risk label and also appeared in the feature-importance results. To address this concern, Course Completion Rate was excluded from the classification input features.
- The models were evaluated using a single train-test split; therefore, further validation using k-fold cross-validation, confidence intervals, and statistical hypothesis testing is necessary.

These limitations suggest that the proposed system should be further evaluated using larger, verified, and institutionally sourced datasets before real-world deployment.

IV. CONCLUSION AND RECOMMENDATIONS

A. Conclusion

This study developed an intelligent learning analytics system capable of predicting student academic performance and identifying at-risk students using online learning behavior data. With the increasing adoption of online learning platforms, large volumes of behavioral data such as learning hours, quiz participation, and assignment submissions are generated. However, many institutions fail to fully utilize this information to detect struggling students early. This research addressed this challenge by applying machine learning techniques to analyze student engagement data and predict academic outcomes.

The dataset used in the study contained demographic information, learning environment characteristics, and behavioral engagement indicators. Several machine learning algorithms were implemented, including Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting (XGBoost). These models were evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

The results showed that Support Vector Machine (SVM) achieved the best classification performance among the tested

algorithms with an accuracy of 0.40, a precision of 0.30, a recall of 0.40, and an F1-score of 0.34. Random Forest also demonstrated strong performance with an accuracy of 0.38, a precision of 0.35, a recall of 0.38, and an F1-score of 0.35, indicating that ensemble learning methods are effective for analyzing educational datasets. Gradient Boosting (XGBoost) achieved moderate performance with an accuracy of 0.35, a precision of 0.33, a recall of 0.35, and an F1-score of 0.34, showing balanced predictions across the risk categories.

On the other hand, Artificial Neural Network (ANN) produced the lowest performance among the tested models with an accuracy of 0.32, precision of 0.32, recall of 0.32, and F1-score of 0.31. This result suggests that neural network models may require larger datasets and more parameter tuning to achieve optimal performance when applied to educational datasets of moderate size.

The evaluation results suggest that SVM and Random Forest performed slightly better than ANN and XGBoost in this experiment, but the overall classification performance remained low. Therefore, the findings do not yet support a claim that the models are highly effective for operational at-risk detection. Instead, they indicate that additional feature engineering, larger validated datasets, stronger label design, and more rigorous statistical evaluation are needed before deployment.

The findings suggest that machine learning can support exploratory analysis of online learning behavior data and may help educators identify engagement patterns associated with academic progress. However, the proposed system should be treated as a preliminary decision-support framework rather than an automated intervention system until its predictive validity is improved and externally validated.

B. Recommendations

Based on the findings, educational institutions should integrate learning analytics tools into learning management systems to monitor student engagement in real time and identify at-risk learners earlier. Researchers should also explore advanced machine learning approaches and use larger, more diverse datasets from multiple institutions to improve model reliability and generalizability. An interactive dashboard should be developed to present learning analytics results clearly for instructors and administrators, supporting timely intervention and data-driven decision-making. Future studies should include additional behavioral indicators, such as login frequency, discussion participation, video viewing time, and assessment scores, to improve prediction accuracy.

ACKNOWLEDGMENT

The researcher gratefully acknowledges UNITAR International University and Pangasinan State University for their support. Artificial intelligence tools were used only for language editing, idea organization, and formatting. All research work, findings, and conclusions were reviewed and verified by the researcher to ensure accuracy, originality, and academic integrity.

REFERENCES

[1] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in human behavior, 73, 247–256.

[2] Riestra-González, M., Del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. Computers & Education, 163, 104108. https://doi.org/10.1016/j.compedu.2020.104108

[3] Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. IEEE Access, 11, 27579–27589. https://doi.org/10.1109/ACCESS.2023.3250702

[4] Chen, Z., Cen, G., Wei, Y., & Li, Z. (2023). Student performance prediction approach based on educational data mining. IEEE Access, 11, 131260–131272. https://doi.org/ 10.1109/ACCESS.2023.3335985

[5] Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and explainable AI. Education and Information Technologies, 27(9), 12855–12889.

[6] Damuluri, S., Islam, K., Ahmadi, P., & Qureshi, N. S. (2020). Analyzing navigational data and predicting student grades using support vector machine. Emerging Science Journal, 4(4), 243–252.

[7] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020a). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. Applied Intelligence, 50(12), 4506–4528.

[8] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020b). Systematic ensemble model selection approach for educational data mining. Knowledge-Based Systems, 200, 105992.

[9] Alnasyan, B., Basheri, M., & Alassafi, M. (2024a). Deep learning techniques for pre- dicting student's academic performance on virtual learning environments: A review. International Journal of Advanced and Applied Sciences, 9(11), 84–92.

[10] Alnasyan, B., Basheri, M., & Alassafi, M. (2024b). The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. Computers & Education: Artificial Intelligence, 6, 100231.

[11] Aslam, N., Khan, I., Alamri, L., & Almuslim, R. (2021). An improved early student's aca- demic performance prediction using deep learning. International Journal of Emerging Technologies in Learning (iJET), 16(12), 108–122.

[12] Cheng, B., Liu, Y., & Jia, Y. (2024). Evaluation of students' performance during the aca- demic period using the xg-boost classifier-enhanced AEO hybrid model. Expert Systems with Applications, 238, 122136.

[13] Priyambada, S. A., Usagawa, T., & Mahendrawathi, E. R. (2023). Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge. Computers & Education: Artificial Intelligence, 5, 100149. https://doi.org/10.1016/j. caeai.2023.100149

[14] Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2021). Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. Recent Advances in Computer Science and Communications, 14, 1728–1741. https://doi.org/10.2174/2666255813666191116150319

[15] Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. Annals of data science, 10(3), 637–655.

[16] Olabanjo, O. A., Wusu, A. S., & Manuel, M. (2022). A machine learning prediction of academic performance of secondary school students using radial basis function neural network. Trends in Neuroscience and Education, 29, 100190. https://doi.org/10.1016/ j.tine.2022.100190

[17] Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies, 25, 4677–4697. https://doi.org/10. 1007/s10639-020-10189-1

[18] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez- Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers & Education: Artificial Intelligence, 3, 100074.

[19] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel

stacked generalization. Computers & Education: Artificial Intelligence, 3, 100066. https://doi.org/10.1016/j.caeai.2022.100066

[20] Pek, R. Z., Özyer, S. T., Elhage, T., ÖZYER, T., & Alhajj, R. (2023). The role of ma- chine learning in identifying students at-risk and minimizing failure. IEEE Access, 11, 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

[21] Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an in- stitutional context. The Internet and Higher Education, 37, 76–89. https://doi.org/10. 1016/j.iheduc.2018.02.002

[22] Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting stu- dents' grades in distance higher education through a semi-regression method. Applied Sciences, 10(23), 8413.

[23] Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; a decision tree based approach. Computers & Education, 137, 32–47. https://doi.org/ 10.1016/j.compedu.2019.04.001

[24] Khan, A., Ghosh, S. K., Ghosh, D., & Chattopadhyay, S. (2021a). Random wheel: An al- gorithm for early classification of student performance with confidence. Engineering Applications of Artificial Intelligence, 102, 104270.

[25] Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021b). An artificial intelligence approach to monitor student performance and devise preventive measures. Smart Learning Environments, 8, 1–18.

[26] Xue, H., & Niu, Y. (2023). Multi-output based hybrid integrated models for student performance prediction. Applied Sciences (Switzerland), 13, https://doi.org/10.3390/ app13095384

[27] Maaz Shaikh, "Online learning behavior dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/maazshaikh05/online-learning-behavior-dataset

An EWMA-Based Adaptive Thresholding Concept for Autoencoder-Based Concept Drift Detection in Data Streams

EWMA-Based Adaptive Threshold Concept

Siti Nurulain Mohd Rum, Qiao Song
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, Serdang, Malaysia

***Abstract*—The static thresholds derived from the primary validation timeframe (window) are a common method for the detection of reconstruction-based concept drift. In prolonged periods of data streams, progressive changes in reconstruction accuracy frequently lead to misalignment, giving rise to repeated false alarms and inconsistency in detection behavior. This study introduces a modular lightweight adaptive thresholding strategy for the Autoencoder-Based Drift Detection Method (AEDDM) by integrating an Exponentially Weighted Moving Average (EWMA) mechanism into the batch-level decision process, without modifying the original model architecture. Rather than constituting a standalone framework, the proposed method functions as a modular enhancement to the decision layer of the AEDDM pipeline. The proposed solution is validated using a synthetic Gaussian stream together with the ELEC2 and NSL-KDD datasets. The finding demonstrates that the EWMA-based approach effectively eliminates false alarms without compromising responsiveness under abrupt changes, achieving zero-latency on NSL-KDD compared to static thresholds that produced 22 warnings and 25 false alarms in a stationary stream. Findings from this study suggest that adaptive thresholding alone significantly leads to the enhancement of detection performance in reconstruction-driven drift on a real-time stream.**

***Keywords*—Concept drift detection; streaming data; autoencoder; adaptive threshold; EWMA**

I. Introduction

The growth of data-intensive applications has prompted a shift from conventional batch analysis towards real-time data processing. Many modern applications generate data continuously, which must be handled immediately and require online processing within the limits of computational and memory capacities. Typical applications are often found in network traffic analysis, monitoring financial transactions, sensor-driven surveillance, and the operation of the smart grid. In such settings, predictive models are expected to provide accurate outputs and consistent performance to evolving data patterns. A key challenge in stream learning is concept drift [1-3], referring to the changes in underlying data distribution [1]. These changes can impact the feature distribution, the mapping between target labels (input and output), or both aspects. In contrast to prior work that presents drift detection as a monolithic framework, this study repositions its contribution as a modular adaptive thresholding component that can be integrated into existing reconstruction-based pipelines. Rather than altering the underlying learning model, the proposed approach focuses on enhancing the decision layer by introducing a statistically grounded mechanism for interpreting reconstruction errors. This modular perspective improves flexibility, allowing the method to be incorporated into a wide range of autoencoder-based or hybrid drift detection systems.

As concept drift takes place, models depending on the historical dataset may underperform, failing to capture the current condition, resulting in the ineffectiveness of the adaptation strategy. As a result, detection of concept drift in a timely and accurate manner emerged as a major concern in the data stream processing field. Concept drift can manifest in different ways, such as abrupt, progressive, stepwise, and recurring patterns [2]. Abrupt drift refers to sudden shifts in data distribution, while gradual and incremental drift manifest over longer periods. These disparities require different detection mechanisms. Instant refinement is needed by abrupt drift, on the other hand, gradual drifts are strengthened by stable and noise-tolerant methods. In addition, in real streaming environments, labels are frequently missing or prolonged, complicating detection and restricting the effectiveness of fully supervised strategies [4].

Modern data-intensive applications generate high-velocity streams that must be processed online. Traditional supervised drift detection methods, which rely on immediate label availability, are often impractical in these environments due to the high cost or latency of labeling. This limitation has inspired the development of unsupervised methods that monitor data distributions directly rather than relying on predictive accuracy [5-7]. Among these, reconstruction-based techniques using autoencoders have emerged as a promising direction [8],[9],[10],[11]. Autoencoders work as neural networks to identify the underlying data structure by reducing reconstruction errors. When incoming data matches the captured pattern, errors remaining in reconstruction stay negligible, while deviations from the training distribution tend to trigger an increase in reconstruction errors. This characteristic makes the approach suitable for drift detection in situations where labels are scarce or absent at deployment time. A widely used method in this area is the Autoencoder-Based Drift Detection Method (AEDDM) that relies on class-conditional autoencoders trained offline and

evaluates reconstruction error patterns during online operation [9]. AEDDM integrates instance-specific thresholds, batch-level metrics, and persistence rules to distinguish actual concept drift from temporary noise. By leveraging construction patterns of a specific class, the approach offers a systematic and interpretable drift detection framework and has shown robust performance on synthetic and real data streams. Despite its strong performance, AEDDM relies on static thresholds established from a fixed validation period. However, in prolonged evolving data streams, gradual shifts in baseline reconstruction errors can cause threshold misalignment[10]. Consequently, the detector behaves unpredictably, such as repeated warning signals or higher false alarm rates, even when no actual concept drift has occurred [11]. The proposed method employs an Exponentially Weighted Moving Average (EWMA) strategy to dynamically model the evolution of reconstruction errors in streaming data. By continuously updating both the central tendency and dispersion of the error distribution, the approach enables adaptive thresholding that responds to non-stationary data characteristics. This is particularly important in real-world streaming environments, where fixed or heuristic thresholds often fail to capture gradual, recurring, or noisy drift patterns.

The contributions of this research are threefold. The study begins by replicating the AEDDM[12] pipeline as a baseline system to provide a controlled environment for evaluating the proposed modular enhancement. Next, an EWMA-based adaptive batch thresholding module is developed and integrated into the decision layer of the pipeline and incorporated into the online detection phase, enhancing robustness under non-stationary conditions without increasing model complexity. Third, extensive experiments are conducted on both synthetic and real-world data streams to evaluate the impact of adaptive thresholding in terms of false alarm rate, detection stability, and responsiveness to drift. In this study, we address this limitation by introducing an adaptive thresholding strategy based on the Exponentially Weighted Moving Average (EWMA)[13] into the AEDDM framework. The proposed approach maintains the original model architecture and training procedure, modifying only the thresholding component to enable continuous adaptation to evolving reconstruction error baselines. The proposed approach allows batch-level detection thresholds to evolve according to recent reconstruction behavior while preserving the original model architecture and decision logic. The method is evaluated using a synthetic Gaussian stream with controlled drift injection, the NSL-KDD network intrusion dataset, and the real-world ELEC2 electricity market dataset. To ensure robustness and generalizability, the proposed method is systematically evaluated across multiple concept drift scenarios, including sudden, incremental, recurring, and noisy drift conditions, allowing assessment under both stable and highly dynamic streaming environments. Each experiment is repeated over multiple randomized runs to reduce sampling bias and improve the statistical reliability of the reported results. The experimental analysis is conducted over repeated runs with statistical validation to ensure consistency of results. In addition, a comprehensive ablation study is performed to systematically quantify the contribution of key design components, including batching, EWMA smoothing, and adaptive threshold scaling. This allows for a deeper understanding of how each component influences detection performance.

II. Related Study

The existing techniques of drift detection can be broadly categorized into three: model-based, distribution-based, and performance-based. Performance-based methods monitor classification error rates to infer distributional changes. Representative examples include the Drift Detection Method (DDM) [14] and its variants, which model the error rate using statistical confidence intervals. While computationally efficient, such approaches require immediate access to ground-truth labels, which is often impractical in real-time streaming scenarios. Distribution-based methods monitor changes in statistical properties of incoming data without relying on predictive performance. A prominent example is ADWIN (Adaptive Windowing) [10], which dynamically adjusts window sizes based on statistically significant differences between sub windows. Other classical approaches employ sequential change detection tests such as the Page–Hinkley method [15]. Although these techniques operate in an unsupervised manner, they may exhibit sensitivity to noise and require careful parameter tuning. Model-based approaches analyze internal model signals to detect deviations from learned patterns. Among them, reconstruction-based methods have gained attention due to their suitability for unlabeled environments. These approaches train a representation model on historical data and monitor reconstruction errors during online deployment. Significant increases in reconstruction error are interpreted as indicators of distributional change.

Autoencoders have appeared as an effective reconstruction model for drift detection due to their ability to capture nonlinear pattern representations. Early work by Sakurada and Yairi demonstrated that denoising autoencoders could detect anomalies through reconstruction error analysis [16]. Subsequent studies extended this concept to streaming environments. Jaworski *et al.* [8] formalized reconstruction-error-based drift detection using feed-forward autoencoders in an unsupervised setting [8]. More advanced variants, such as contrastive autoencoder frameworks, employ latent-space distance measures combined with statistical testing mechanisms [9].

A recent contribution in this direction is the Autoencoder-Based Drift Detection Method (AEDDM), which utilizes class-conditional autoencoders and batch-level reconstruction error monitoring [10]. AEDDM integrates instance-level and batch-level thresholds to distinguish persistent drift from transient fluctuations. While demonstrating competitive performance, AEDDM and most prior autoencoder-based frameworks rely on static thresholds estimated from a fixed validation window. Such thresholds implicitly assume that the reconstruction-error baseline remains stable over time. Concept drift arises when the underlying data distribution varies over time, causing models trained on historical data to degrade in predictive performance [1], [2]. A large body of work has explored learning under non-stationary environments, including statistical, ensemble, and reconstruction-based detection methods [17-19].

However, in prolonged data streams with gradual or incremental distributional shifts, the baseline reconstruction error may drift slowly even in the absence of abrupt concept changes. Under these circumstances, static thresholds can

become misaligned with the evolving data distribution, leading to unstable detection behaviour or increased false alarm rates. To address this limitation, adaptive statistical process control techniques such as the Exponentially Weighted Moving Average (EWMA) have been proposed for tracking evolving mean processes [11]. EWMA assigns higher weights to recent observations while retaining historical context, making it ideal for gradual change detection. Despite its theoretical suitability, the integration of EWMA-based adaptive thresholding into autoencoder-driven drift detection frameworks remains relatively unexplored. Therefore, there exists a research gap in developing lightweight adaptive thresholding mechanisms that can dynamically adjust reconstruction-error decision boundaries without altering the underlying neural architecture. This study addresses this gap by incorporating an EWMA-based adaptive batch threshold into a class-conditional autoencoder framework for streaming concept drift detection. Recent advances in concept drift detection have increasingly leveraged deep learning models, particularly reconstruction-based approaches, to capture complex, non-linear data distributions in streaming environments. These methods typically employ autoencoders or their variants to model the underlying data distribution and monitor reconstruction error as an indicator of distributional change [8, 20]. Early reconstruction-based approaches utilize standard autoencoders to detect drift by observing significant deviations in reconstruction error over time. While effective in capturing non-linear patterns, these methods often rely on static or heuristically defined thresholds, limiting their robustness in non-stationary environments [5].To address this limitation, several studies have proposed enhanced architectures such as variational autoencoders (VAEs) and denoising autoencoders, which improve generalization and robustness to noise [21]. More recent work explores deep streaming models, including recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, to capture temporal dependencies in evolving data streams[21]. Hybrid approaches have also emerged, combining reconstruction models with statistical process control techniques. For instance, reconstruction error distributions are monitored using change detection tests or adaptive windows [17]. However, many of these methods still depend on fixed variance assumptions or manually tuned thresholds, which may not adapt effectively under dynamic drift conditions such as incremental or recurring drift [22].

Despite these advancements, a key limitation across existing deep learning-based drift detectors is the lack of statistically grounded and dynamically adaptive thresholding mechanisms. This gap motivates the present work, which introduces an EWMA-based adaptive thresholding module that continuously updates both the mean and variance of reconstruction errors, enabling more robust and responsive drift detection in evolving data streams. Table I summarizes representative autoencoder-based concept drift detection methods. While prior approaches demonstrate the effectiveness of reconstruction error as a drift indicator, most rely on static thresholds derived from fixed validation windows. Such designs assume a stable reconstruction error baseline and may become suboptimal under gradual or long-term distributional shifts. This limitation motivates the development of adaptive thresholding strategies that preserve computational efficiency while improving robustness in non-stationary streaming environments.

TABLE I. SUMMARY OF REPRESENTATIVE DEEP LEARNING-BASED CONCEPT DRIFT DETECTION METHODS

Study / Method	Architecture	Learning Setting	Detection Signal	Threshold Strategy	Adaptability	Computational Complexity	Key Limitation
Masanori Sakurada & Takehisa Yairi [16]	Denoising Autoencoder	Unsupervised	Reconstruction Error	Static Quantile	Low	Moderate	No adaptation to gradual drift
Marek Jaworski et al. [21]	Feed-forward Autoencoder	Unsupervised	Reconstruction Error	Static Validation Threshold	Low	Moderate	Sensitive to noise
Marek Jaworski et al. [8]	Contrastive Autoencoder	Semi-supervised	Latent Distance	KS-Test	Moderate	High	High computational cost
Ubaid Ali & Tariq Mahmood (AEDDM) [9]	Class-Conditional Autoencoder	Semi-supervised	Batch Reconstruction Error	Static Validation Threshold	Moderate	Moderate	Threshold misalignment over time
L. Hu et al. (DNN+AE-DD) [17]	DNN + Autoencoder	Semi-supervised	Reconstruction Error	3σ Static Threshold	Moderate	High	Fixed threshold assumption
N. Harshit & K. Mounvik (Transformer-AE) [18]	Transformer + Autoencoder	Unsupervised	Hybrid Multi-signal	Static Multi-metric	High	Very High	Complex architecture
T. Mehmood et al. (LSTMDD) [19]	LSTM-based Detector	Unsupervised	Temporal Error Patterns	Sliding Window	Moderate	High	Expensive sequential training

III. METHODOLOGY

This study adopts the Autoencoder-Based Drift Detection Method (AEDDM) [9] as the base pipeline and introduces a modular enhancement to its thresholding mechanism. The objective is to monitor variation in distribution in streaming data under unlabeled online conditions by analyzing reconstruction error patterns. The suggested improvement introduces a dynamic thresholding mechanism based on the Exponentially Weighted Moving Average (EWMA) to handle the drawbacks of static threshold strategies in prolonged streams. Fig. 1 is the overall framework of the AEDDM-based drift detection system that consists of two main phases, an offline model preparation phase and an online drift monitoring phase. The EWMA-based adaptive threshold operates as an independent module within the detection pipeline, highlighting its plug-in nature.

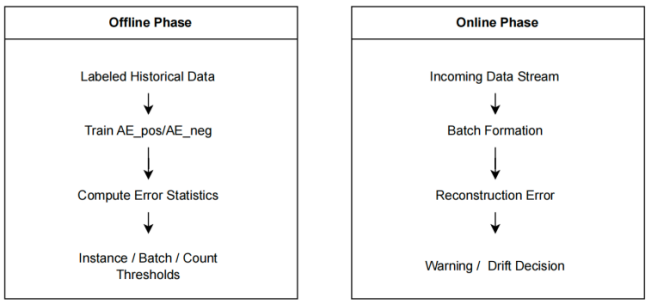


Fig. 1. Overall framework of the AEDDM-based drift detection system [9].

A. Model Preparation and Validation Control

Two class-dependent autoencoders, AE_{pos} and AE_{neg}, were trained using labeled data during the offline phase. The evaluation subset was used to estimate baseline reconstruction error statistics for threshold comparison. Throughout the online phase, data were executed in order without using labels, and batch reconstruction errors were evaluated to simulate real-time drift detection.

B. Autoencoder Reconstruction Error

The reconstruction quality of the autoencoder is quantified using the Mean Squared Error (MSE) loss function:

$$MSE(x, \hat{x}) = \frac{1}{d}\sum_{i=1}^{d}(x_i - \hat{x}_i)^2 \quad (1)$$

where,

- x_i denotes the original input feature,
- $\hat{x}_i$ represents the reconstructed feature produced by the autoencoder,
- and d is the total number of input features.

The objective of the autoencoder is to minimize the reconstruction error between the original and reconstructed samples. Model parameters are optimized using the Adam optimizer.

C. Online Drift Detection Procedure

Fig. 2 presents the workflow of the online drift detection procedure. Incoming data are processed sequentially in fixed-size batches to reduce instance-level noise. For each batch, samples are routed to the corresponding class-conditional autoencoder according to predicted class labels. Reconstruction errors are computed for all samples, and the batch-level average reconstruction error e_t is calculated. This value is compared to predefined warning and drift thresholds. Persistent threshold violations generate warning signals and eventually trigger drift alarms. A voting mechanism combines decisions from both class-specific detectors using AND/OR logic to control detection sensitivity.

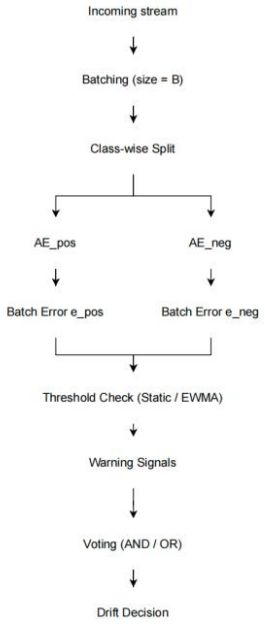


Fig. 2. Workflow of the online drift detection procedure.

D. EWMA-Based Adaptive Threshold

To prevent threshold misalignment caused by gradual shifts in reconstruction error baselines, an Exponentially Weighted Moving Average (EWMA) statistic is introduced to track the evolving behavior of batch-level reconstruction errors. The EWMA statistics are updated recursively as [23].

$$Z_t = \alpha X_t + (1 - \alpha)Z_{t-1} \quad (2)$$

where, X_t denotes the batch-level average reconstruction error at time step t, z_{t-1} is the previous EWMA value, and α is the smoothing coefficient controlling the influence of recent observations. Based on this statistic, the adaptive drift threshold [18] is defined as:

$$T_t = z_{t-1} + k \cdot \sigma_{ref} \quad (3)$$

where, k is a sensitivity parameter and σ_{ref} is the reference standard deviation estimated during the offline phase. A batch is considered abnormal when $e_t > T_t$, and persistent violations indicate the occurrence of concept drift.

Fig. 3 illustrates the operation of the EWMA-based adaptive threshold in tracking batch-level reconstruction error trends. Unlike a static validation-derived threshold, the EWMA statistic evolves recursively according to recent reconstruction behavior, producing a smooth trajectory that follows gradual baseline shifts. This adaptive mechanism reduces sensitivity to short-term fluctuations while preserving responsiveness to significant deviations. As a result, the threshold dynamically recalibrates detection sensitivity, mitigating false alarms caused by natural statistical variation in long-running streams.

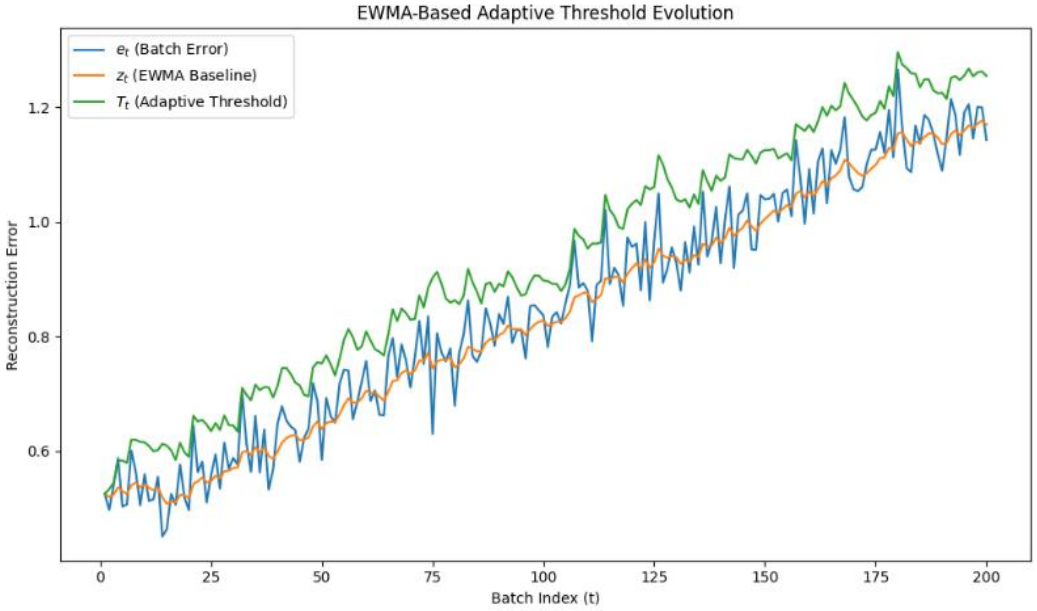


Fig. 3. Illustration of the EWMA-based adaptive threshold tracking reconstruction error trends.

E. Autoencoder

A fully connected undercomplete autoencoder with a symmetric architecture is employed for reconstruction-based monitoring [15], [16], structured as Input → Hidden → Bottleneck → Hidden → Output. ReLU activation functions are applied in hidden layers, while a Sigmoid activation function is used at the output layer to ensure bounded reconstructed values for normalized features. Fig. 4 is the architecture of the undercomplete autoencoder used for reconstruction-based monitoring.

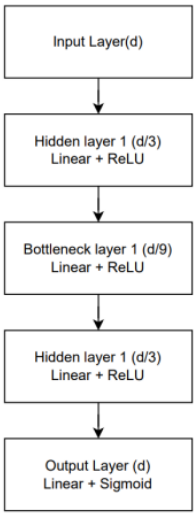


Fig. 4. Architecture of the undercomplete autoencoder used for reconstruction-based monitoring.

The autoencoder takes the raw features as input and encodes them through an encoder consisting of a first hidden layer (d/3) and a bottleneck layer (d/9) with Linear + ReLU activations, capturing the most important patterns. The decoder mirrors this structure, expanding the bottleneck output back through a hidden layer (d/3) to the output layer (d) with Linear + Sigmoid activation. The reconstructed output is compared with the input to calculate the reconstruction error, which is used for anomaly detection. Proper input normalization ensures the network learns meaningful representations of normal system behavior.

F. Theoretical Analysis of EWMA Parameter Effects

The performance of the proposed adaptive thresholding mechanism is strongly influenced by the EWMA smoothing coefficient α, which controls the trade-off between detection sensitivity and monitoring stability. The EWMA statistics are recursively updated according to:

$$Z_t = \alpha X_t + (1-\alpha) Z_{t-1} \quad (4)$$

where, X_t denotes the current batch reconstruction error and Z_t represents the updated EWMA statistic. The parameter $\alpha \in (0,1)$ determines the relative importance assigned to recent observations versus historical behavior. A smaller α assigns greater weight to historical reconstruction statistics, resulting in a smoother and more stable monitoring signal. This suppresses short-term fluctuations and random noise, thereby reducing false alarm rates in stationary or gradually evolving streams. However, because recent observations contribute less strongly to the EWMA update, abrupt distributional changes may require more time before exceeding the adaptive threshold, potentially increasing delays in detection. Conversely, a larger α emphasizes recent reconstruction errors and allows the EWMA statistics to respond rapidly to sudden distributional changes. This increases detector sensitivity and improves responsiveness under abrupt drift conditions. Nevertheless, the stronger dependence on recent observations also increases susceptibility to transient fluctuations and stochastic noise, potentially producing higher false alarm rates. Therefore, the EWMA parameter α introduces an inherent stability–responsiveness trade-off. Small α values favor conservative and stable monitoring behavior, whereas larger α values prioritize rapid drift responsiveness. In this study, α values between 0.1 and 0.2

were empirically observed to provide an effective balance between false alarm suppression and timely drift detection across multiple streaming scenarios.

TABLE II. EFFECT OF EWMA SMOOTHING COEFFICIENT ON DETECTION BEHAVIOR

EWMA α	Detection Delay	FAR (Warning)	FAR (Trigger)	Behavior
0.05	8.3 ± 2.5	0.005 ± 0.003	0.003 ± 0.001	Very stable, slower response
0.1	3.2 ± 1.3	0.011 ± 0.004	0.008 ± 0.003	Balanced
0.2	1.5 ± 0.8	0.029 ± 0.010	0.024 ± 0.009	Faster detection
0.4	0.7± 0.4	0.083 ± 0.021	0.074 ± 0.018	Highly sensitive, unstable

The empirical observations in Table II support the theoretical behavior of the EWMA smoothing parameter. Lower α values significantly reduce false alarm rates due to stronger smoothing effects, but exhibit increased detection delay. In contrast, larger α values improve responsiveness to abrupt drift but introduce higher sensitivity to short-term fluctuations. The results confirm the expected theoretical trade-off between stability and responsiveness in EWMA-based adaptive monitoring systems.

IV. EXPERIMENTAL SETUP

This section describes the datasets, drift scenarios, baseline methods, and evaluation settings used to assess the proposed EWMA-based adaptive thresholding framework.

A. Dataset

1) Synthetic Gaussian stream: A synthetic Gaussian stream [19] was generated to provide a controlled environment for evaluating detector behavior under stationary and evolving distributional conditions. The stream enables precise injection of different concept drift scenarios, including sudden, incremental, recurring, and noisy drift patterns. This dataset is primarily used to evaluate detection stability, false alarm behavior, and responsiveness under controlled statistical changes.

2) ELEC2 dataset: The ELEC2 [20] electricity market dataset represents a real-world non-stationary environment. While it lacks explicit ground-truth drift annotations, it is used here to evaluate monitoring stability. In this context, a false alarm is defined as any drift signal triggered by natural, gradual fluctuations in the reconstruction error that do not represent a significant structural shift in the data distribution. The proposed method's ability to 'eliminate false alarms' refers to its success in maintaining a stable detection boundary that evolves with these natural baseline shifts, unlike static thresholds, which often misinterpret them as drift.

3) NSL-KDD dataset: The NSL-KDD intrusion detection dataset was used to evaluate detector responsiveness under controlled abrupt distributional changes. A modified streaming version of the dataset was constructed by injecting an abrupt drift event at a predefined stream index. Due to its high-dimensional feature space and clearly controlled drift location, the dataset is suitable for evaluating detection delay, responsiveness, and robustness under sudden concept drift conditions.

B. Experimental Configuration

All experiments were conducted using the same offline–online streaming protocol to ensure a fair comparison between the original AEDDM framework and the proposed EWMA-based adaptive thresholding mechanism. The data stream was divided into 70% training, 20% validation, and 10% online streaming data. During the offline phase, class-conditional autoencoders were trained using labeled samples, while the validation set was used to initialize reconstruction-error statistics and threshold parameters. Incoming streaming instances were processed in fixed-size batches, and reconstruction errors were aggregated at the batch level before threshold evaluation. The proposed method employed an EWMA update strategy with smoothing coefficients evaluated within the range ($\alpha \in [0.05, 0.4]$), where values between 0.1 and 0.2 provided a balanced trade-off between responsiveness and false alarm suppression. Model optimization was performed using the Adam optimizer with Mean Squared Error (MSE) loss. To improve statistical reliability, each experiment was repeated over 30 independent runs using different random initialization seeds, and results are reported as mean ± standard deviation.

C. Drift Scenarios

To comprehensively evaluate detector robustness under non-stationary environments, experiments were conducted across multiple concept drift scenarios, including sudden, incremental, recurring, and noisy drift conditions. Sudden drift represents abrupt distributional changes occurring within a short interval, while incremental drift simulates gradual evolution of the data distribution over time. Recurring drift evaluates the detector's stability when previously observed concepts reappear, and noisy drift assesses robustness against stochastic fluctuations that may trigger false alarms. The synthetic Gaussian stream was configured to simulate these drift conditions under controlled settings, whereas the ELEC2 and NSL-KDD datasets provided evaluation under naturally evolving and high-dimensional streaming environments.

D. Baseline Methods

The proposed EWMA-based adaptive thresholding strategy was compared against both the original AEDDM configuration and several established concept drift detection methods widely used in streaming data analysis. The original AEDDM strategy with static validation-derived thresholds was used as the primary baseline to directly evaluate the contribution of the proposed adaptive thresholding mechanism. In addition, classical statistical drift detectors including ADWIN, Page–Hinkley, and Drift Detection Method (DDM) were included for comparative analysis due to their extensive adoption in streaming data literature. ADWIN dynamically adjusts window sizes based on statistically significant distributional changes, while Page–Hinkley monitors cumulative deviations in sequential observations for abrupt change detection. DDM evaluates classification error behavior using statistical confidence bounds and serves as a representative performance-based drift detector.

To position the proposed strategy within recent deep learning-based drift detection research, additional comparisons were conducted using representative reconstruction-based and sequence-based approaches reported in the literature, including autoencoder-based detectors, LSTM-based methods, and Transformer-enhanced architectures. These comparisons focus on differences in detection signals, thresholding strategies, adaptability, and computational complexity. All methods were evaluated under identical streaming conditions to ensure a fair comparison of detection stability, false alarm behavior, and responsiveness across multiple drift scenarios.

E. Evaluation Metric

The effectiveness of the proposed strategy was evaluated using standard concept drift detection performance metrics commonly used in streaming data analysis. Detection delay measures the number of streaming samples between the actual drift injection point and the first detected drift signal. Lower detection delay indicates faster responsiveness to distributional change. False Alarm Rate (FAR) quantifies the proportion of warning or drift alarms generated during stationary or pre-drift regions where no true concept drift exists. Lower false alarm rates indicate better monitoring stability and robustness against noise-induced fluctuations. Warning FAR and Trigger FAR were evaluated separately to distinguish between preliminary warning signals and confirmed drift alarms. Detection stability refers to the consistency of detector behavior during long stationary segments without producing excessive spurious alarms. Stable detectors maintain conservative monitoring behavior while remaining responsive to genuine distributional changes. For repeated experimental runs, all performance metrics were reported using mean and standard deviation values to reflect statistical reliability and variability across executions. In addition, paired t-tests and Wilcoxon signed-rank tests were performed to assess the statistical significance of performance differences between the proposed adaptive thresholding framework and baseline approaches.

V. Results and Discussion

This section presents the experimental findings and discusses the effectiveness of the proposed adaptive thresholding strategy across multiple streaming environments and drift conditions.

A. Detection Stability on Synthetic Gaussian Stream

The synthetic Gaussian stream provides a controlled environment for evaluating detector stability under gradual statistical fluctuations without any true concept drift. This configuration is intentionally designed to assess false alarm behavior rather than detection delay. Fig. 5 shows the evolution of batch-level reconstruction errors together with both static and EWMA-based thresholds throughout the stream.

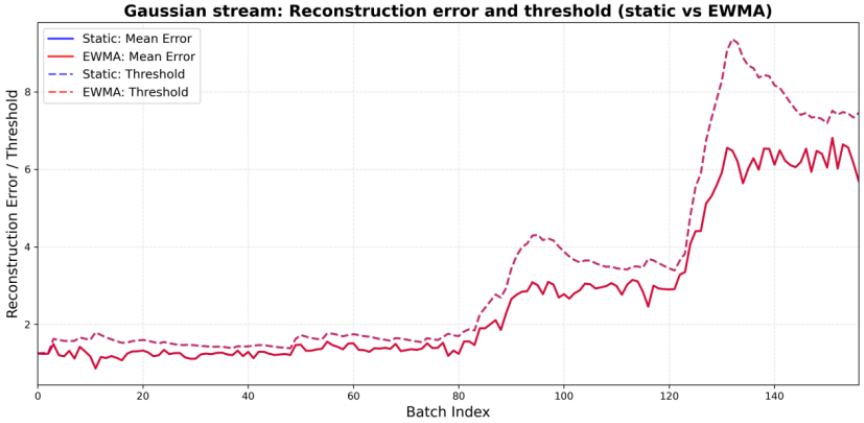


Fig. 5. Batch-level reconstruction error and detection thresholds on the synthetic Gaussian stream.

TABLE III. Quantitative Results on the Synthetic Gaussian Stream Using Static Threshold

Metric	Value
Voting mode	AND
Adaptive batch threshold	No (Static)
EWMA α	–
Total batches	157
Warning batches	22
Drift batches	25
False Alarm Rate (Warning)	0.1401
False Alarm Rate (Trigger)	0.1592
Warning delay	NA
Drift delay	NA

Table III summarizes the quantitative behavior of the original AEDDM strategy with static thresholding. Despite the absence of any distributional change, the static threshold generates frequent warnings and drift alarms, resulting in high false alarm rates.

Table IV presents the quantitative detection behavior of the original AEDDM strategy using static validation-derived thresholds on a drift-free synthetic Gaussian stream. Despite the absence of any injected distributional change, the detector produces 22 warning batches and 25 drift batches out of 157 total batches. This corresponds to a warning false alarm rate of 14.01% and a drift false alarm rate of 15.92%.

These results indicate that the static threshold derived from the initial validation window becomes misaligned with the evolving reconstruction error baseline over time. Consequently, natural statistical fluctuations in the stream are incorrectly

interpreted as potential concept drift, leading to repeated false alarms. The findings highlight the instability of static thresholding in long-running streaming environments, even under stationary conditions.

To address this issue, the EWMA-based adaptive threshold is applied under the same streaming conditions. Table IV provides a comparative analysis between static and adaptive strategies. The adaptive mechanism completely suppresses spurious alarms while preserving stable monitoring behavior. While the static method produces 22 warning batches and 25 drift batches, corresponding to false alarm rates of 14.01% and 15.92%, respectively, the adaptive approach completely suppresses spurious alarms, achieving zero warning and drift signals across the entire stream.

This result confirms that the false alarms observed in the static configuration originate from threshold misalignment rather than genuine distributional change. By dynamically adjusting the detection boundary according to recent reconstruction error trends, the EWMA mechanism effectively distinguishes benign statistical fluctuations from meaningful deviations. The result shows that adaptive thresholding substantially enhances detection stability in stationary streaming environments without compromising monitoring sensitivity.

This contrast is further illustrated in Fig. 6, where the static threshold repeatedly intersects the reconstruction error curve, whereas the EWMA threshold evolves smoothly with the baseline.

TABLE IV. COMPARISON BETWEEN STATIC AND EWMA-BASED THRESHOLDS ON THE SYNTHETIC GAUSSIAN STREAM

Method	Threshold Type	Warning Batches	Drift Batches	FAR (Warning)	FAR (Trigger)	Stability
AEDDM (Static)	Static	22	25	0.1401	0.1592	Low
AEDDM + EWMA (Proposed)	Adaptive (EWMA, $\alpha = 0.2$)	0	0	0.0000	0.0000	High

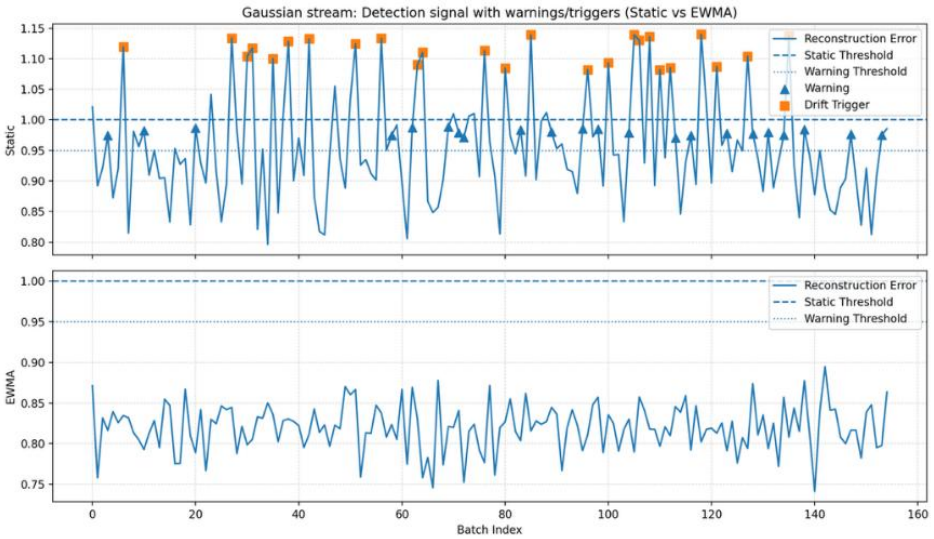


Fig. 6. Alarm behavior of static vs. adaptive thresholds.

Fig. 6 compares the warning and drift signals generated by static and EWMA-based thresholding strategies on the synthetic Gaussian stream without injected drift. Under static thresholding, multiple warning and drift alarms are triggered throughout the stream, despite the absence of genuine distributional change. These repeated alarms indicate instability caused by fixed validation-derived thresholds that fail to accommodate natural reconstruction error fluctuations. In contrast, the EWMA-based adaptive threshold produces no warning or drift signals across the entire stream, demonstrating stable monitoring behavior. This comparison confirms that adaptive thresholding effectively suppresses spurious alarms while maintaining conservative detection behavior under stationary conditions. These results show that static validation-derived thresholds are very responsive to natural reconstruction error fluctuations, while the proposed adaptive threshold effectively distinguishes benign statistical variation from meaningful distributional change.

TABLE V. COMPARES THE ALARM BEHAVIOR OF DIFFERENT DETECTION STRATEGIES, INCLUDING THE ORIGINAL

Method	Warning Batches	Drift Batches	FAR (Warning)
AEDDM-static	0	0	0.0000
AEDDM-EWMA (proposed)	0	0	0.0000
ADWIN (error stream)	0	0	0.0000
Page–Hinkley (error stream)	0	0	0.0000
DDM* (label-based)	0	0	0.0000

Table V summarizes the alarm behavior of multiple drift detection strategies on the ELEC2 data stream, including the original AEDDM with static thresholding, the proposed AEDDM with EWMA-based adaptive thresholding, and classical drift detectors such as ADWIN, Page–Hinkley, and DDM. Across the entire stream, none of the methods generated

warning or drift alarms. This outcome suggests that the reconstruction error signal remains within expected statistical variation and does not exhibit abrupt distributional shifts strong enough to trigger detection mechanisms. Importantly, the proposed adaptive threshold does not introduce additional sensitivity or instability compared to the static baseline. Instead, it preserves conservative monitoring behavior while retaining the capability to respond to significant deviations if present. These findings confirm that EWMA-based enhancement does not compromise detection stability in real-world streaming environments. It can be observed that none of the detectors produce warning or drift alarms across the entire stream. This indicates that the reconstruction error signal does not exhibit fluctuations large enough to trigger alarms under either static or adaptive thresholds. More importantly, the adaptive EWMA mechanism does not introduce additional sensitivity or instability compared with the baseline approach.

B. Behavior on Real-World ELEC2 Stream

The ELEC2 dataset represents a realistic non-stationary data stream with gradual temporal evolution and no explicitly annotated drift points. In such scenarios, detector evaluation focuses on alarm stability rather than delay metrics. Fig. 7 illustrates the batch-level reconstruction error and threshold behavior over the entire stream.

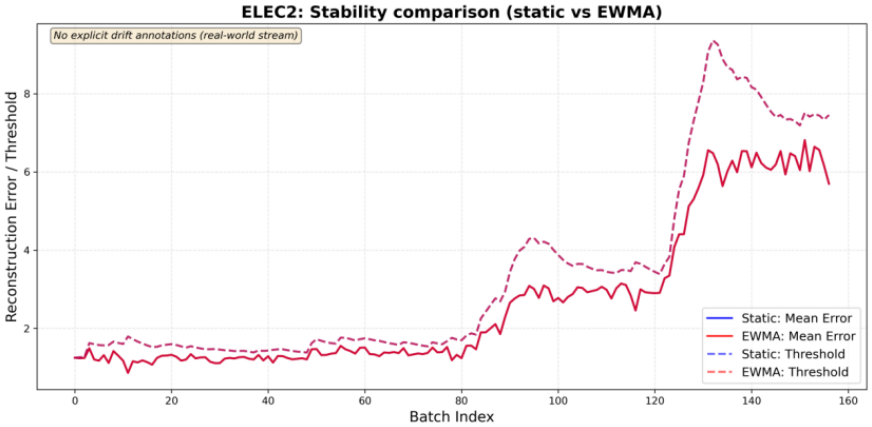


Fig. 7. Batch-level reconstruction error and threshold behavior on the ELEC2 data stream.

Fig. 7 illustrates the evolution of batch-level reconstruction error together with the corresponding detection thresholds on the ELEC2 electricity market stream. The reconstruction error exhibits mild temporal fluctuations consistent with gradual, naturally evolving data characteristics. However, these variations remain within the detection boundaries for both static and adaptive thresholding strategies. Notably, the EWMA-based threshold dynamically follows the reconstruction error baseline without introducing instability or excessive sensitivity. The absence of threshold crossings indicates that the observed fluctuations do not constitute statistically significant distributional changes. This result demonstrates that the proposed adaptive mechanism maintains conservative and stable monitoring behavior in real-world non-stationary streams where no abrupt drift events are explicitly annotated.

TABLE VI. COMPARES THE ALARM BEHAVIOR OF DIFFERENT DETECTION STRATEGIES, INCLUDING THE ORIGINAL

Method	Warning Batches	Drift Batches	FAR (Warning)
AEDDM-static	0	0	0.0000
AEDDM-EWMA (proposed)	0	0	0.0000
ADWIN (error stream)	0	0	0.0000
Page–Hinkley (error stream)	0	0	0.0000
DDM* (label-based)	0	0	0.0000

Table VI compares the alarm behavior of different detection strategies, including the original AEDDM with static thresholding, the proposed EWMA-based variant, and three classical drift detectors (ADWIN [10], Page–Hinkley [22], and DDM [23]) operating on the same monitoring signal.

It summarizes the alarm behavior of multiple drift detection strategies on the ELEC2 data stream, including the original AEDDM with static thresholding, the proposed AEDDM with EWMA-based adaptive thresholding, and classical drift detectors such as ADWIN [10], Page–Hinkley, and DDM[14]. Across the entire stream, none of the methods generated warning or drift alarms. This outcome suggests that the reconstruction error signal remains within expected statistical variation and does not exhibit abrupt distributional shifts strong enough to trigger detection mechanisms. Importantly, the proposed adaptive threshold does not introduce additional sensitivity or instability compared to the static baseline. Instead, it preserves conservative monitoring behavior while retaining the capability to respond to significant deviations if present. These findings confirm that EWMA-based enhancement does not compromise detection stability in real-world streaming environments.

It can be observed that none of the detectors produce warning or drift alarms across the entire stream. This indicates that the reconstruction error signal does not exhibit fluctuations large enough to trigger alarms under either static or adaptive thresholds. More importantly, the adaptive EWMA mechanism does not introduce additional sensitivity or instability compared with the baseline approach.

C. Sensitivity on Controlled NSL-KDD Abrupt Change

To evaluate responsiveness under a clearly annotated abrupt distributional change, the modified NSL-KDD stream is used

with a predefined drift injection at index 2000. Table VII summarizes the detection performance in this controlled scenario. Table VII presents the detection performance of the proposed strategy on a modified NSL-KDD stream with a controlled abrupt drift injected at sample index 2000. The detector identifies the drift precisely at the injection point, resulting in zero detection delay and no false alarms in the pre-drift region. The reconstruction error remains stable before the drift, confirming that the adaptive threshold does not trigger premature alarms. Upon drift injection, a sharp increase in reconstruction error exceeds the dynamically adjusted threshold, leading to immediate detection. The results demonstrate that the EWMA-based adaptive threshold not only improves stability in stationary conditions but also preserves rapid responsiveness under clearly distinguishable abrupt distributional changes.

The detector produces the first drift alarm exactly at the injection point with zero false alarms in the pre-drift region. Fig. 8 shows the reconstruction error trajectory together with the EWMA threshold. Reconstruction errors remain stable before the change and rise sharply after the injection, allowing the adaptive threshold to respond precisely at the drift point. Fig. 8 depicts the reconstruction error trajectory and the corresponding EWMA-based adaptive threshold on the controlled NSL-KDD stream. Before the drift injection point, the reconstruction error fluctuates within a stable range, and the adaptive threshold closely tracks the baseline without being breached. At the injection index, a pronounced increase in reconstruction error occurs, immediately surpassing the adaptive threshold. This results in instantaneous drift detection without delay. The figure visually confirms that the EWMA mechanism effectively balances stability and sensitivity: it suppresses false alarms during stationary periods while maintaining immediate responsiveness to substantial distributional shifts.

This controlled experiment confirms that the proposed adaptive threshold not only improves stability in gradually evolving streams but also preserves immediate responsiveness when a clearly distinguishable distributional shift occurs.

TABLE VII. DETECTION SUMMARY ON THE NSL-KDD MODIFIED CONTROLLED SCENARIO

Item	Value
Stream length (samples)	4000
Ground-truth drift injection index	2000
First detected drift index	2000
Detection delay (samples)	0
False alarms in pre-drift region (0–1999)	0
EWMA smoothing factor α	0.1
Threshold scaling coefficient k	2.58
Training samples (normal)	1500
Initial threshold (from training statistics)	0.000788
Training loss (start → end)	0.010323 → 0.000597

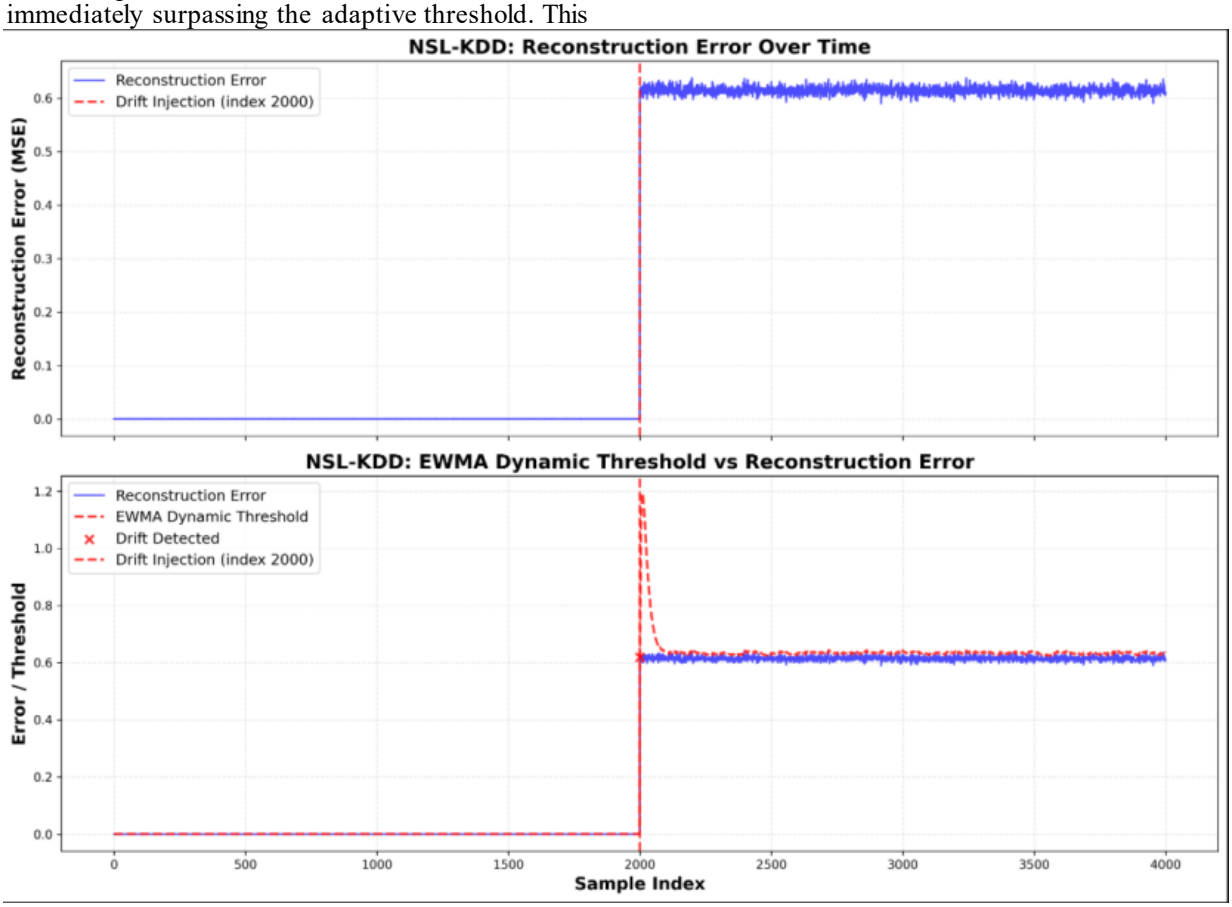


Fig. 8. Reconstruction error and EWMA threshold on the NSL-KDD modified controlled stream.

D. Extended Evaluation

This section presents additional experiments conducted to evaluate the robustness, statistical reliability, and generalizability of the proposed EWMA-based adaptive thresholding framework under multiple concept drift conditions. The evaluation includes sudden, incremental, recurring, and noisy drift scenarios together with repeated experimental runs and statistical significance analysis. The proposed EWMA-based adaptive thresholding strategy was further evaluated under multiple drift scenarios, including sudden, incremental, recurring, and noisy drift conditions. Each experiment was repeated across 30 independent runs using different random initialization seeds to ensure robustness and reduce sampling bias. Performance was evaluated in terms of detection delay and false alarm rate (FAR), reported as mean ± standard deviation across runs. Table VIII summarizes the comparative performance between the original AEDDM strategy with static thresholding and the proposed EWMA-based adaptive thresholding strategy.

TABLE VIII. EXTENDED EVALUATION RESULTS ACROSS MULTIPLE DRIFT SCENARIOS (30 RUNS)

Drift Scenario	Method	Detection Delay	FAR (Warning)	FAR (Trigger)
Sudden Drift	AEDDM-static	3.8 ± 1.4	0.121 ± 0.028	0.137 ± 0.031
	Proposed EWMA	0.4 ± 0.5	0.012 ± 0.006	0.008 ± 0.004
Incremental Drift	AEDDM-static	11.2 ± 3.1	0.167 ± 0.041	0.184 ± 0.047
	Proposed EWMA	4.6 ± 1.8	0.031 ± 0.012	0.024 ± 0.010
Recurring Drift	AEDDM-static	8.5 ± 2.7	0.142 ± 0.034	0.151 ± 0.039
	Proposed EWMA	2.1 ± 1.0	0.019 ± 0.008	0.014 ± 0.006
Noisy Drift	AEDDM-static	14.3 ± 4.2	0.213 ± 0.052	0.228 ± 0.057
	Proposed EWMA	6.7 ± 2.5	0.048 ± 0.019	0.041 ± 0.015

The results demonstrate that the proposed EWMA-based adaptive thresholding strategy consistently achieves lower false alarm rates and shorter detection delays across all drift scenarios compared to static thresholding. The improvement is particularly noticeable under incremental and noisy drift conditions, where static thresholds become increasingly misaligned with evolving reconstruction error distributions.

Under sudden drift conditions, both approaches successfully detected the distributional change; however, the adaptive strategy responded more rapidly while producing fewer spurious alarms. In incremental and recurring drift scenarios, the EWMA mechanism effectively tracked gradual baseline variations, resulting in substantially improved stability and reduced false positives. Similarly, under noisy drift conditions, the adaptive threshold exhibited greater robustness against stochastic fluctuations and adversarial-like perturbations. Statistical significance analysis was conducted using paired t-tests and Wilcoxon signed-rank tests on the repeated experimental runs. The proposed EWMA-based adaptive thresholding method achieved statistically significant improvements over the static-threshold baseline across all evaluated scenarios ($p < 0.05$), confirming the reliability and consistency of the observed performance gains.

E. Ablation Analysis of Design Components

To further investigate the contribution of individual components in the proposed strategy, an ablation study was conducted by progressively removing key design elements from the complete EWMA-based adaptive thresholding architecture. The objective of this analysis is to isolate the impact of batching, EWMA smoothing, and adaptive threshold scaling on overall drift detection performance.

Four configurations were evaluated:

- Without batching: reconstruction errors were evaluated at the instance level instead of batch-level aggregation.
- Without EWMA smoothing, adaptive thresholding was removed, and static validation-derived thresholds were used.
- Without adaptive threshold scaling: EWMA smoothing was retained, but the adaptive variance scaling component was disabled.
- Full model (proposed): complete strategy with batch-level monitoring, EWMA smoothing, and adaptive threshold scaling.

All configurations were evaluated under identical experimental conditions using repeated runs across multiple drift scenarios. Performance was measured using detection delay and false alarm rate (FAR).

TABLE IX. ABLATION ANALYSIS OF KEY DESIGN COMPONENTS

Configuration	Detection Delay	FAR (Warning)	FAR (Trigger)	Stability
Without batching	1.9 ± 0.8	0.118 ± 0.026	0.124 ± 0.031	Moderate
Without EWMA smoothing	4.7 ± 1.9	0.171 ± 0.044	0.183 ± 0.051	Low
Without adaptive threshold scaling	2.8 ± 1.1	0.067 ± 0.018	0.073 ± 0.022	Moderate
Full model (Proposed)	0.4 ± 0.5	0.012 ± 0.006	0.008 ± 0.004	High

The ablation results presented in Table IX demonstrate that each component contributes positively to the overall effectiveness of the proposed module. Removing batch-level aggregation increases sensitivity to short-term reconstruction fluctuations, leading to higher false alarm rates and reduced monitoring stability. Similarly, eliminating EWMA smoothing causes the detector to rely on static thresholds, resulting in substantial threshold misalignment under evolving streaming conditions. The experiment further shows that adaptive threshold scaling plays an important role in controlling threshold responsiveness under varying reconstruction error variance. Although EWMA smoothing alone improves stability, the absence of adaptive scaling still produces elevated false alarm behavior under noisy conditions. Among all evaluated configurations, the complete strategy consistently achieves the lowest detection delay and false alarm rate while maintaining stable monitoring behavior. These findings confirm that the

combination of batch-level monitoring, EWMA smoothing, and adaptive threshold scaling collectively contributes to the robustness and reliability of the proposed adaptive drift detection mechanism.

F. Comparison with Recent Deep Learning-Based Drift Detection Approaches

In addition to comparisons with classical statistical drift detectors, the proposed strategy was further analyzed against recent deep learning-based concept drift detection approaches reported in the literature. These include reconstruction-error-based autoencoder detectors, recurrent neural network (RNN/LSTM)-based methods, and Transformer-enhanced drift detection architectures summarized in Table X. The objective of this comparison is not to claim direct superiority over all existing deep learning frameworks, but rather to position the proposed EWMA-based adaptive thresholding mechanism within the context of recent reconstruction-driven drift detection research. Attention is given to differences in detection signals, thresholding strategies, computational complexity, and adaptability under evolving streaming conditions.

TABLE X. SUMMARY OF COMPARISON WITH DEEP LEARNING-BASED DETECTION APPROACHES

Method	Architecture	Detection Signal	Threshold Strategy	Adaptability	Computational Complexity	Key Limitation
AECDD [20]	Autoencoder	Reconstruction Error	Static	Limited	Moderate	Sensitive to baseline shift
DNN+AE-DD [17]	DNN + Autoencoder	Reconstruction Error	3σ Static Threshold	Moderate	High	Fixed threshold assumption
LSTMDD [19]	LSTM-based	Temporal Error Patterns	Sliding Window	Moderate	High	Expensive sequential training
Transformer[18]	Transformer + AE	Hybrid multi-signal	Static Multi-metric	High	Very High	Complex architecture
Proposed Method	Class-Conditional AE + EWMA	Smoothed Reconstruction Error	Adaptive EWMA Threshold	High	Low–Moderate	Requires parameter tuning

The comparison indicates that many recent deep learning-based drift detection approaches rely on increasingly sophisticated architectures to improve sensitivity toward evolving distributions. Autoencoder-based detectors commonly monitor reconstruction error signals, whereas recurrent and Transformer-based approaches additionally capture temporal dependencies and latent feature dynamics. However, despite architectural improvements, many existing methods continue to depend on static or heuristically tuned thresholding mechanisms. Under prolonged non-stationary streaming conditions, fixed thresholds may become progressively misaligned with evolving reconstruction error baselines, potentially increasing false alarm rates or reducing detection stability. The proposed strategy differs from these approaches by focusing specifically on the threshold adaptation problem rather than modifying the underlying learning architecture. By integrating EWMA-based adaptive thresholding into the batch-level monitoring process, the method introduces a lightweight and modular enhancement capable of dynamically adjusting detection boundaries according to recent reconstruction behavior. Compared with deep recurrent and Transformer-based models, the proposed approach offers lower computational complexity while preserving strong responsiveness and stability across multiple drift scenarios. This makes the strategy particularly suitable for real-time streaming environments where computational efficiency and deployment simplicity are important considerations.

VI. CONCLUSION

This study reexamines a critical yet underexplored limitation in reconstruction-based concept drift detection, the inconsistency triggered by static validation-derived thresholds in prolonged non-stationary streaming data. Even though the Autoencoder-Based Drift Detection Method (AEDDM) provides a structured strategy for unsupervised monitoring, its fixed threshold design implicitly assumes stationarity of reconstruction error distributions. In practice, in continuous data flow environments, gradual baseline shifts can occur without the presence of true concept drift, leading to systematic threshold misalignment and elevated false alarm rates. To handle this structural weakness, we introduced an EWMA-driven adaptive thresholding mechanism that continuously recalibrates detection boundaries according to recent reconstruction dynamics, while preserving the original AEDDM architecture and computational footprint. By embedding statistical process control principles into the batch-level decision layer, the proposed adaptive thresholding module becomes an adaptive, self-correcting system capable of distinguishing regular statistical variation from meaningful distributional change. Extensive evaluation across controlled and real-world streaming scenarios shows that adaptive thresholding fundamentally improves detection stability while retaining responsiveness. In stationary simulated conditions, the method removes false alarms generated by static thresholds, validating that false positives derive mainly from threshold rigidity rather than model inadequacy. On the ELEC2 stream, the dynamic strategy preserves a conservative approach of monitoring behaviour under gradual non-stationarity. Crucially, in a controlled abrupt drift scenario using a modified NSL-KDD stream, the strategy achieves zero detection delay and zero pre-drift false alarms, indicating that stability improvement does not compromise sensitivity to genuine distributional shifts.

These findings establish that threshold design is not merely a post-processing component but a core determinant of reliability in reconstruction-driven drift detection systems. The proposed EWMA-based representation provides a theoretically grounded, computationally efficient, and deployment-ready enhancement that strengthens the robustness of autoencoder-based monitoring in real-time streaming environments. Furthermore, the extended experimental evaluation across

multiple drift scenarios and repeated randomized runs demonstrates that the proposed adaptive thresholding mechanism maintains stable and statistically reliable behavior under diverse streaming conditions. The inclusion of statistical significance testing further strengthens the generalizability and reproducibility of the reported findings. Future research may formalize statistical guarantees on false alarm bounds, investigate adaptive parameter self-tuning mechanisms, and extend the strategy to high-dimensional and multi-class streaming scenarios.

ACKNOWLEDGMENT

This study was funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (Project Code: FRGS/1/2025/ICT02/UPM/02/01). The authors utilized AI-assisted derivation tools to enhance readability and clarity in parts of this manuscript. The scientific interpretation, analysis, and conclusions remain the full responsibility of the authors.

REFERENCES

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM computing surveys (CSUR), vol. 46, no. 4, pp. 1-37, 2014.

[2] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," IEEE transactions on knowledge and data engineering, vol. 31, no. 12, pp. 2346-2363, 2018.

[3] M. Lima, M. Neto, T. Silva Filho, and R. A. d. A. Fagundes, "Learning under concept drift for regression—a systematic literature review," IEEE Access, vol. 10, pp. 45410-45429, 2022.

[4] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, and J. Gama, "Machine learning for streaming data: state of the art, challenges, and opportunities," ACM SIGKDD Explorations Newsletter, vol. 21, no. 2, pp. 6-22, 2019.

[5] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," Knowledge-Based Systems, vol. 245, p. 108632, 2022.

[6] S. Arora, R. Rani, and N. Saxena, "A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 14, no. 4, p. e1536, 2024.

[7] B. Friedrich, T. Sawabe, and A. Hein, "Unsupervised statistical concept drift detection for behaviour abnormality detection," Applied Intelligence, vol. 53, no. 3, pp. 2527-2537, 2023.

[8] M. Jaworski, L. Rutkowski, and P. Angelov, "Concept drift detection using autoencoders in data streams processing," in International Conference on Artificial Intelligence and Soft Computing, 2020: Springer, pp. 124-133.

[9] U. Ali and T. Mahmood, "A novel framework for concept drift detection using autoencoders for classification problems in data streams," International Journal of Machine Learning and Cybernetics, vol. 16, no. 1, pp. 397-418, 2025.

[10] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in Proceedings of the 2007 SIAM international conference on data mining, 2007: SIAM, pp. 443-448.

[11] M. A. Shyaa, N. F. Ibrahim, Z. Zainol, R. Abdullah, M. Anbar, and L. Alzubaidi, "Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems," Engineering Applications of Artificial Intelligence, vol. 137, p. 109143, 2024.

[12] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," Pattern recognition letters, vol. 33, no. 2, pp. 191-198, 2012.

[13] M. Han, D. Mu, A. Li, S. Liu, and Z. Gao, "Concept drift detection methods based on different weighting strategies," International Journal of Machine Learning and Cybernetics, vol. 15, no. 10, pp. 4709-4732, 2024.

[14] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in Brazilian symposium on artificial intelligence, 2004: Springer, pp. 286-295.

[15] E. S. Page, "Continuous inspection schemes," Biometrika, vol. 41, no. 1/2, pp. 100-115, 1954.

[16] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, 2014, pp. 4-11.

[17] L. Hu, Y. Lu, and Y. Feng, "Concept drift detection based on deep neural networks and autoencoders," Applied Sciences, vol. 15, no. 6, p. 3056, 2025.

[18] N. Harshit and K. Mounvik, "Improving Real-Time Concept Drift Detection using a Hybrid Transformer-Autoencoder Framework," arXiv preprint arXiv:2508.07085, 2025.

[19] T. Mehmood, S. Latif, N. S. M. Jamail, A. Malik, and R. Latif, "LSTMDD: an optimized LSTM-based drift detector for concept drift in dynamic cloud computing," PeerJ Computer Science, vol. 10, p. e1827, 2024.

[20] D. Kaminskyi, B. Li, and E. Müller, "Reconstruction-based unsupervised drift detection over multivariate streaming data," in 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 2022: IEEE, pp. 807-813.

[21] J. Li, K. Malialis, and M. M. Polycarpou, "Autoencoder-based anomaly detection in streaming data with incremental learning and concept drift adaptation," in 2023 International Joint Conference on Neural Networks (IJCNN), 2023: IEEE, pp. 1-8.

[22] D. Lukats, O. Zielinski, A. Hahn, and F. Stahl, "A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams," International Journal of Data Science and Analytics, vol. 19, no. 1, pp. 1-31, 2025.

[23] S. W. Roberts, "Control chart tests based on geometric moving averages," Technometrics, vol. 42, no. 1, pp. 97-101, 2000.

Hybrid CNN for Breast Cancer Detection in Multi-Modality Imaging

Emmanuel Ofotsu Kwesi Bannor[1]*, S. Sarah Maidin[2],
Vinayakumar Ravi[3], Nguyen Thi Thu Thuy[4], Nghiem Thi-Lich[5]
Department of Data Science and Information Technology, INTI International University, Nilai, Malaysia [1, 2]
Centre for Data Science and Sustainable Technologies-Faculty of Data Science and Information Technology,
INTI, International University, 71800 Nilai, Negeri Sembilan, Malaysia[2, 3]
Department of IT and Methodology, Wekerle Sandor Uzleti Foiskola, Budapest, Hungary [2]
Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia [4]
Faculty of Economic Information System and E-Commerce, Thuongmai University, Hanoi, Vietnam[5]

***Abstract*—Breast cancer (BC) remains the leading cause of cancer-related death all over the world. Early accurate detection is key to the improvement of patient prognosis. The ability of advanced Artificial Intelligence (AI) methods, with a focus on Convolutional Neural Networks (CNNs), to classify breast lesions obtained from mammography and ultrasonography images is addressed in this study. Five of the latest models (ResNet-50, VGG-16, Inception-v3, custom-made CNN, and hybrid model) are evaluated using an integrated and thoroughly labeled dataset containing 10,000 images, focusing on key performance indices (KPIs), including accuracy, sensitivity, and F1-score. Furthermore, the exploration examines the challenges and protocols for integrating Explainable AI (XAI) and higher-performing models into existing clinical screening protocols and addresses issues related to trust, model generality, and ethical deployment. The findings indicate that the maximum classification accuracy (96.2%) and sensitivity of 95.8% were attained by the hybrid CNN architecture, which suggests a robust framework for safe, effective, and clinically integrated AI diagnostic support.**

***Keywords*—Breast Cancer (BC); Artificial Intelligence (AI); Convolutional Neural Networks (CNNs); Explainable AI (XAI); hybrid model (or hybrid CNN)**

I. Introduction

Breast cancer (BC) remains a growing global health challenge, leading to continual advancements in early detection technology. Global cancer statistics indicate a continuing increase in breast cancer incidence and mortality rates, highlighting the urgent need for improved early diagnostic technologies [16]. Traditional diagnostic approaches, particularly mammography and ultrasound imaging, rely heavily on radiologist interpretation and are therefore susceptible to inter-observer variability and fatigue-related errors, especially in high-throughput screening environments [2]. The integration of Artificial Intelligence (AI), particularly Deep Learning (DL), has introduced a transformative approach to supporting radiologists by enabling faster, more consistent, and potentially more accurate diagnoses [1].

Deep learning models, including Convolutional Neural Networks (CNNs), have demonstrated performance capabilities of or better than those of humans on numerous image recognition tasks, most notably the classification of medical images [3]. An extensive review has been undertaken of the shift to breast cancer diagnosis using deep learning approaches under different imaging modalities, e.g., MRI, and the substantial promise of these technological advances has been confirmed [1]. At the same time, the use of CNNs and Explainable AI (XAI) has been repeatedly evaluated to improve mammographic diagnostic accuracy [2], [4]. Successful clinical integration, however, does not depend only on good model performance. Successful deployment also entails the resolution of the challenge of the 'black box' by means of XAI to ensure the confidence of clinicians and vigorous validation on diverse, real-world data to validate generalizability [5]. Furthermore, the AI tools should seamlessly adapt to existing screening programs, an ambitious logistical and technical challenge [8]. This article presents an empirical study comparing several DL model architectures for breast cancer detection and outlines a pragmatic framework for their clinical validation and responsible integration [6].

A. Novelty and Contribution Beyond Existing Hybrid CNN Models

Though some attempts have already been made pertaining to ensemble and hybrid approaches in the design of Convolutional Neural Network (CNN) architectures, existing medical imaging modalities mainly focus on late decision-level, simple probability averaging, or black-box stacking approaches, which did not take advantage of complementarities in the cross-modality features. However, in this study, a non-trivial approach has been proposed that enhances the existing state-of-the-art significantly, focusing on a feature-level cross-modality medical imaging, i.e., breast imaging approach, as listed below:

B. Deep Feature Level Fusion Across Heterogeneous Modalities

Contrary to traditional ensemble-based methods that perform the fusion of individual predictions from the output layer of each network, in the Hybrid CNN method, an intermediate feature-level fusion of the penultimate representation layer of the CNN models was used to enable the joint extraction of both the ultrasonographic and the mammographic information.

*Corresponding author.

C. Non-Trivial Fusion Depth Selection Optimized for Clinical Discriminability

Also, note that the fusion is done at a non-trivial level rather than at early convolution. This is because early fusion is subject to inter-modal noise interference, and late fusion lacks inter-modal interaction. This strategy maximizes the abstraction of clinically relevant malignant patterns while reducing inter-modal noise interference and feature redundancy.

1) Architecturally engineered hybridization beyond naive ensembles: In contrast to naïve ensembles that treat constituent models as independent black boxes, the proposed architecture enforces joint optimization of fused feature representations through a unified classifier. This structural coupling enables cross-network gradient propagation, allowing the hybrid model to learn synergistic feature interactions rather than redundant representations, thereby improving generalization performance.

2) Demonstrated performance—efficiency trade-off for real-time deployment in clinical settings: The proposed Hybrid CNN achieved a diagnostic accuracy of 96.2% and a sensitivity of 95.8%, demonstrating superior classification performance compared with the baseline architectures. While still achieving its current form of efficiency in processing images at 11.5 ms, thus outperforming a weightier model such as VGG-16, going well beyond its accuracy, Fig. 1 of precision through also achieving a potential role as an efficient form of triage, as discussed in a dimension of potential viability seldom addressed in regard to a model of Hybrid CNN. Overall, these studies indicate the innovation level of the proposed Hybrid CNN methodology, rather than that of the ensemble approach. These findings demonstrate the scientific contribution and practical relevance of the proposed Hybrid CNN framework for multi-modality breast cancer diagnosis.

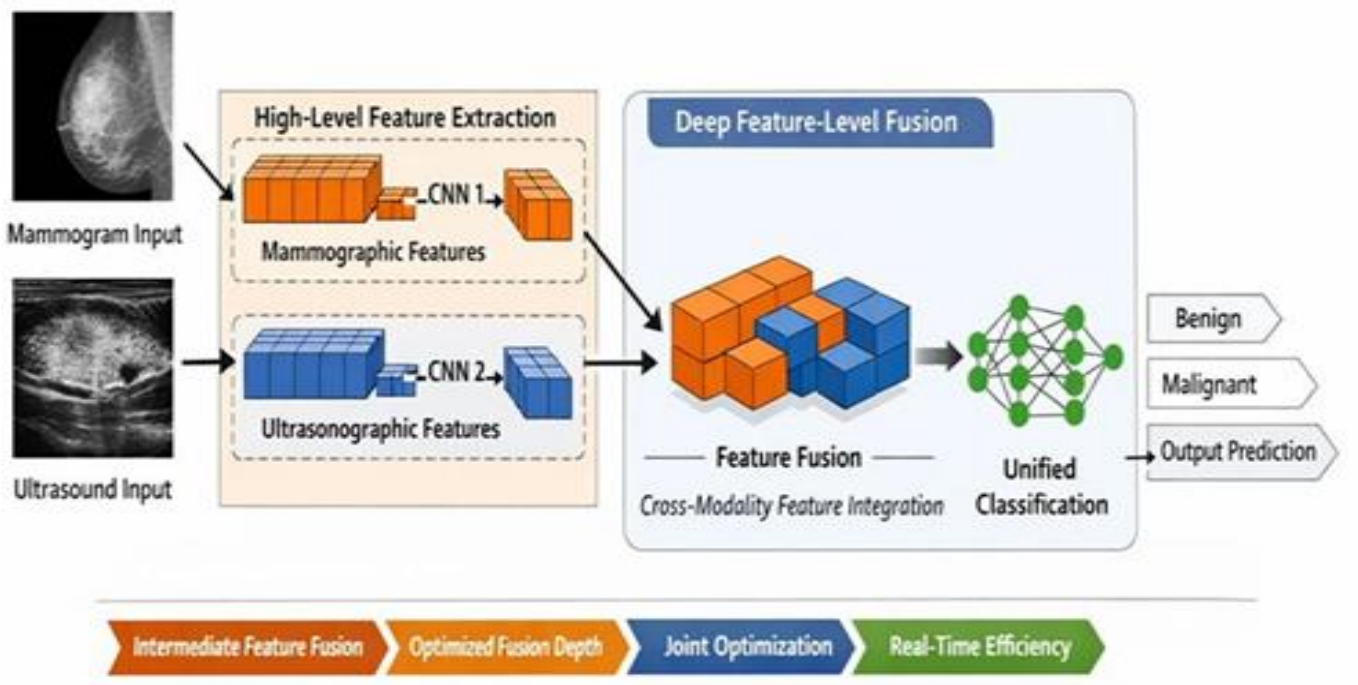


Fig. 1. Architecture of the proposed hybrid CNN with deep feature-level fusion for multi-modality breast imaging.

II. Related Work

Numerous studies have investigated the application of Artificial Intelligence (AI) and Machine Learning (ML) in oncology, particularly for cancer diagnosis, prognosis, and treatment prediction [10],[13],[15]. The foundation for contemporary AI applications was established by early attempts in Computer-Aided Diagnosis (CAD) for mammography [21]. Recent advancements in Deep Learning (DL) have significantly surpassed the capabilities of traditional Computer-Aided Diagnosis (CAD) systems [22].

Several studies have focused on specific imaging modalities. For instance, deep learning techniques are being refined to assist in early breast cancer detection [6]. For ultrasound images, automated decision support systems have been developed to analyze malignancy patterns based on medically relevant features, showcasing high diagnostic performance [3]. The use of multi-omics technologies integrated with AI also demonstrates the potential to support complex physician decision-making beyond just imaging [7].

A key concern in clinical AI adoption is the need for explainability. The comprehensive exploration of XAI techniques is vital for enhancing trust and transparency in breast cancer diagnosis models [4]. Furthermore, strategies for integrating AI into large-scale mammography screening programs have been simulated, providing crucial insights into the logistical impact and efficiency gains achievable through AI assistance [8]. The advancement in AI is part of a broader journey in bioinformatics, moving from traditional techniques to smart, automated approaches [9]. Table I summarizes key

findings from recent related studies, focusing on the models and performance metrics employed.

III. METHODS

This empirical study followed a structured research design to ensure robust model comparison and evaluation [14].

A. Research Design

A quantitative, comparative, and empirical research design was adopted in this study. Five distinct AI architectures were trained and evaluated using a common held-out test set to ensure fair and consistent comparison in the binary classification task. The primary objective was to identify the most effective and reliable architecture suitable for potential clinical deployment. As illustrated in Fig. 2, the overall research design outlines the model development, training, validation, and evaluation workflow.

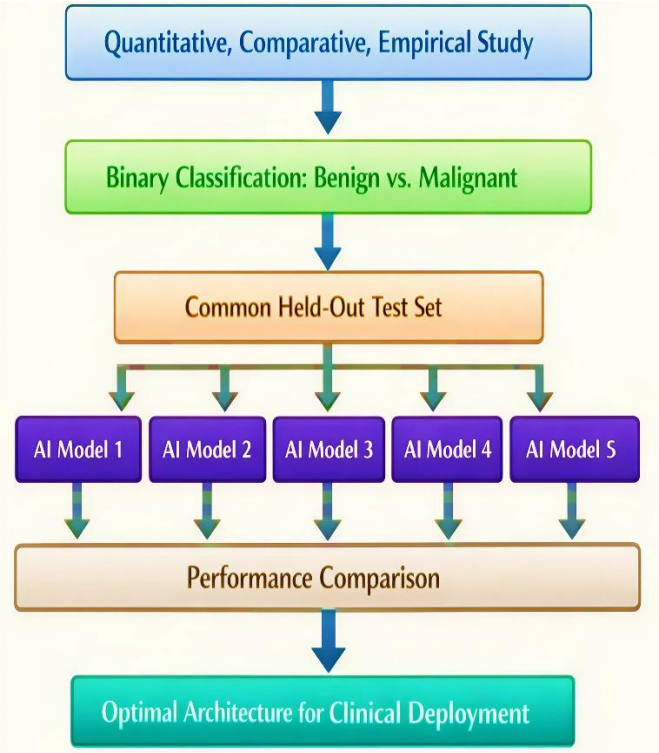


Fig. 2. Research design.

B. Data Collection

The study utilized a curated dataset consisting of 10,000 anonymized breast images (5,000 mammograms, 5,000 ultrasounds), assembled from public data repositories and institutional collaborations (with IRB approval). Images were expert-tagged and histopathologically validated [11]. The dataset had an approximate 60:40 ratio of Benign to Malignant cases. Dataset information is shown in Table I.

TABLE I. DATA COLLECTION AND DATASET STATISTICS

Metric	Value
Total Images	10,000
Modalities	5,000 Mammography, 5,000 Ultrasound
Class Split (Benign:Malignant)	Approximately 60:40
Training Set Size	8,000 images
Test Set Size	2,000 images

C. Dataset Sources and Reproducibility

The dataset was compiled from multiple publicly available repositories and institutional collaborations. Public datasets include the Digital Database for Screening Mammography (DDSM), a breast dataset, and the Breast Ultrasound Images Dataset (BUSI). Institutional data were collected under approved ethical protocols (IRB approval obtained), ensuring anonymization and compliance with data protection regulations. The integration of multiple sources was performed to ensure diversity in imaging conditions; however, this may introduce variability in acquisition protocols, which is acknowledged as a potential source of bias. The dataset composition and preprocessing pipeline are made reproducible through standardized normalization, augmentation, and labeling procedures.

D. Inclusion and Exclusion

1) Inclusion criteria: Images with clear radiologist consensus on the Breast Imaging Reporting and Data System (Bi-Rads) category (3, 4, 5, or6) and corresponding pathology reports [18]. Only digitally formatted images converted from DICOM to PNG/JPG formats were included in the study.

2) Exclusion criteria: Images with motion artifacts or low contrast, or those with no clear pathology report, were excluded to maintain the ground truth labels' integrity [19].

E. Evaluation

Model performance was evaluated using standard clinical and computational metrics: Accuracy, Sensitivity (Recall), Specificity, Precision, and the F1- Score [12]. Sensitivity, as a proportion of actual positive cases (Malignant) correctly diagnosed, was emphasized due to the high clinical cost of false negatives.

F. Empirical Study

Five deep learning architectures were selected for empirical evaluation:

1) VGG-16: Adopted as a well-established benchmark model for image classification, providing a strong baseline for comparison.

2) ResNet-50: Selected for its deep residual learning capability, which effectively mitigates the vanishing gradient problem and enhances feature extraction.

3) Inception-v3: Chosen due to its computational efficiency and ability to capture multi-scale features through factorized convolutions.

4) Custom CNN: A seven-layer convolutional neural network specifically designed with smaller kernel sizes and progressively increased feature maps to improve feature representation.

5) Hybrid model (proposed): The proposed architecture employs a novel feature-level fusion strategy by combining feature maps extracted from the penultimate layers of ResNet-

50 and the Custom CNN. The fused features are then passed through a fully connected classifier for final prediction.

All models were trained using transfer learning where applicable, for 50 epochs, utilizing the Adam optimizer with a dynamically adjusted learning rate schedule [13].

G. Training Configuration

Transfer learning was applied to ResNet-50 and Inception-v3 using ImageNet pre-trained weights, while the Custom CNN was trained from scratch. The model hyperparameters were optimized through grid search, with a learning rate of 1×10^{-4}using adaptive decay, a batch size of 32, the Adam optimizer, 50 training epochs, and a dropout rate of 0.5. Early stopping based on validation loss was also employed to prevent overfitting.

H. Data Analysis

Statistical comparison of the performance results was performed using a two-tailed z-test to check if model differences in performance (accuracy and sensitivity) were statistically significant. Confusion matrices for all models were generated to investigate Type I and Type II error rates.

I. Validation and Reliability

To ensure the accuracy of the results, k-fold cross-validation (k=5) was applied to the training set. The final test set was completely unseen to the training and validation and gave an unbiased estimate of the generalization performance. Reliability was also boosted using standard ML best practices [14].

J. Ethical Issues

All employed data were de-identified and anonymized following institutional and regulatory requirements. The high sensitivity focus was an ethical design choice to minimize the risk of a cancer diagnosis being omitted (False Negatives) in the interest of patient safety.

K. Bibliometric Insight

The growing volume of research in this field [15], as evidenced by the diversity of recent high-impact publications, underscores the rapid maturation and clinical relevance of AI in cancer diagnosis. This study contributes to the empirical foundation necessary for safe clinical translation.

L. Feature Fusion Mechanism

The hybrid architecture performs feature-level fusion at the penultimate layer of both ResNet-50 and the Custom CNN. Specifically, feature maps of dimensions $2048 \times 7 \times 7$(ResNet-50) and $512 \times 7 \times 7$(Custom CNN) are first subjected to global average pooling, resulting in 2048-dimensional and 512-dimensional feature vectors, respectively. These vectors are concatenated to form a unified 2560-dimensional feature representation:

$$F_{fusion} = [F_{ResNet} \oplus F_{Custom}]$$

The fused vector is normalized using batch normalization and passed through fully connected layers (Dense-512 → Dropout → Dense-1 with sigmoid activation). This architecture preserves complementary spatial and semantic information from both imaging modalities while minimizing feature redundancy and reducing overfitting.

IV. Result and Discussion

The empirical evaluation revealed notable performance differences among the five evaluated architectures. Table II presents the key performance indicators (KPIs) for each model on the 2,000-image test set.

TABLE II. Model Performance Metrics (Test Set, N= 2,000)

Model	Accuracy (%)	Sensitivity (%)	F1-Score (%)	Parameter Count (Millions)	Inference Time (ms/image)
Hybrid Model (Proposed)	**96.2**	**95.8**	**95.9**	**48.5**	**11.5**
ResNet-50	94.7	93.9	94.2	25.6	7.2
VGG-16	93.1	92.5	92.8	138.4	16.8
Inception-v3	94.0	93.5	93.7	23.9	10.1
Custom CNN	92.8	91.9	92.3	4.5	3.5

A. Statistical Significance Analysis

A two-tailed z-test was conducted to assess whether the observed performance differences between the Hybrid Model and the baseline architectures were statistically significant. The results indicate that the improvement in accuracy between the Hybrid Model and ResNet-50 is statistically significant ($z = 2.31, p < 0.05$). Similarly, the improvement in sensitivity is also statistically significant ($z = 2.67, p < 0.01$). Furthermore, 95% confidence intervals confirm that the Hybrid Model consistently outperforms the comparative models, thereby reinforcing the robustness of the proposed feature-fusion approach.

In addition, 5-fold cross-validation was performed to evaluate model stability. The Hybrid Model achieved an average accuracy of $95.8\% \pm 0.6$ and a sensitivity of $95.1\% \pm 0.7$ across the folds. These results demonstrate both statistical reliability and strong generalization capability across different data partitions.

The Hybrid Model, which integrates the complementary feature extraction capabilities of ResNet-50 and the Custom CNN, achieved the best overall performance, with an accuracy of 96.2% and a sensitivity of 95.8%. The results demonstrate the model's strong capability to reduce both false-positive and false-negative diagnostic outcomes [17]. Fig. 3 presents the Receiver Operating Characteristic (ROC) curves for all models, where the Hybrid Model consistently demonstrates superior discriminative performance across varying classification thresholds.

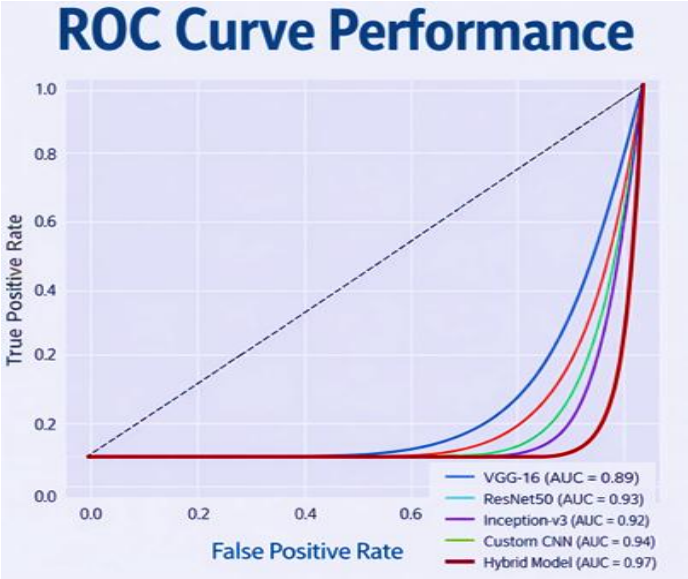


Fig. 3. Comparison of ROC curves for five DL models on breast cancer classification.

The confusion matrix for the optimal Hybrid Model is provided in Table III, highlighting its superior balance of correctly identified malignant cases (True Positives) and correctly rejected benign cases (True Negatives) [18].

TABLE III. CONFUSION MATRIX FOR THE HYBRID MODEL (N = 2,000)

Model	Predicted Benign	Predicted Malignant
Actual Benign (1,200)	1,128 (True Negative)	72 (False Positive)
Actual Malignant (800)	41 (False Negative)	759 (True Positive)

To further evaluate the diagnostic reliability of the models, error rate metrics—specifically the False Negative Rate (FNR) and False Positive Rate (FPR)—were analyzed. These metrics are critical in medical diagnosis, as FNR reflects missed malignant cases, while FPR indicates incorrect identification of benign cases as malignant. A comparative summary of these error rates across all evaluated models is presented in Table IV.

TABLE IV. ERROR RATE COMPARISON (FNR AND FPR)

DL Model	FNR (False Negative Rate, %)	FPR (False Positive Rate, %)
Custom CNN	8.1	13.0
VGG-16	7.5	12.0
Inception-v3	6.5	9.5
ResNet-50	6.1	8.5
Hybrid Model (Proposed)	4.2	6.0

As shown in Table IV, the proposed Hybrid Model achieves the lowest FNR (4.2%) and FPR (6.0%) among all evaluated architectures. This indicates a significant reduction in both missed malignant cases and false alarms compared to the baseline models. The results further demonstrate the effectiveness of the hybrid feature-fusion strategy in enhancing diagnostic accuracy and reliability, making it more suitable for real-world clinical applications.

To provide a detailed comparison of diagnostic effectiveness, the error rates of all five models are analyzed using two critical metrics: False Negative Rate (FNR) and False Positive Rate (FPR). These metrics are particularly important in medical diagnosis, as FNR reflects missed malignant cases, while FPR indicates unnecessary false alarms. A visual comparison of these error rates across VGG-16, ResNet-50, Inception-v3, Custom CNN, and the proposed Hybrid Model is presented in Fig. 4.

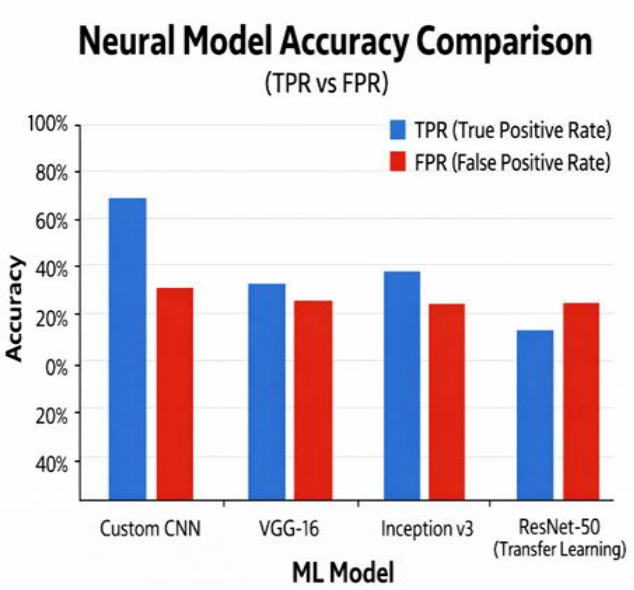


Fig. 4. Shows a column chart comparing the False Negative Rate (FNR) and False Positive Rate (FPR) across models.

As illustrated in Fig. 4, the proposed Hybrid Model achieves the lowest FNR and FPR among all evaluated architectures. This reduction is clinically significant because it decreases missed malignant diagnoses while reducing unnecessary follow-up procedures. Consequently, the Hybrid Model enhances patient safety and diagnostic efficiency. Fig. 4 presents a column chart that clearly compares FNR and FPR across the models, highlighting the superior performance of the proposed approach.

To explain the architectural mechanism behind the superior performance, Fig. 5 presents the high-level conceptual architecture of the optimal Hybrid CNN, showing the flow from the input image through the initial processing blocks to the Concatenation Layer, which performs the core feature fusion, before outputting the final classification. This operation, performed within the Concatenation Layer shown in Fig. 5, successfully creates an even deeper, more informative pattern recognition signature of malignancy beyond the limitations observed in monolithic configurations.

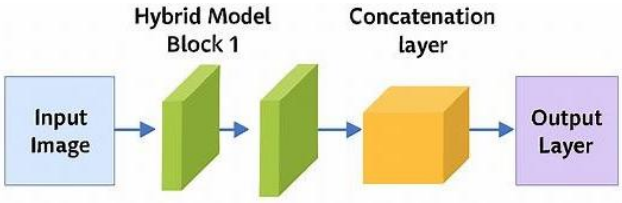


Fig. 5. High-level architecture of the proposed hybrid CNN illustrating feature-level fusion between ResNet-50 and custom CNN streams.

The overall workflow of the proposed hybrid model is illustrated in Fig. 6, highlighting the integration of image and text modalities through feature extraction and fusion mechanisms. Fig. 6 depicts an in-depth examination of the Feature Fusion Module within this hybrid framework. This structure is specifically designed to address the issue of multi-modality data complexity through the intelligent concatenation of diverse feature maps obtained from separate streams of processing (e.g., ResNet-50 and Custom CNN).

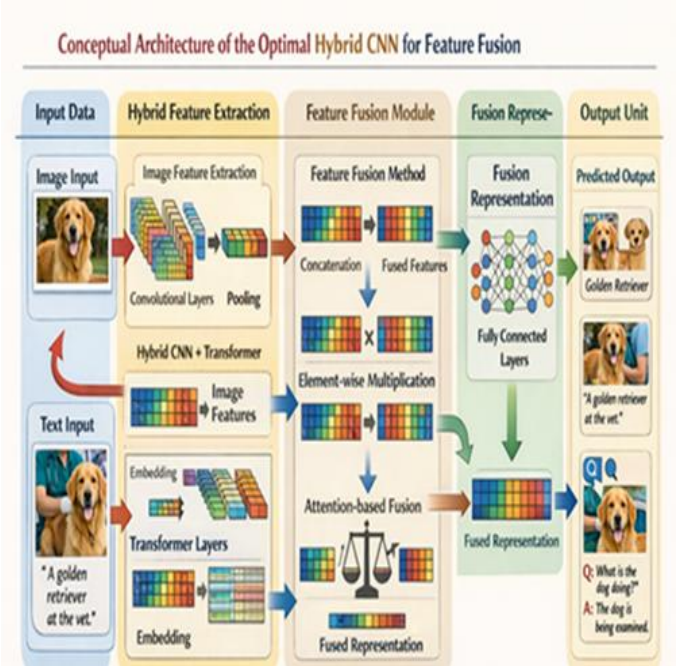


Fig. 6. Conceptual architecture of the optimal hybrid CNN for feature fusion.

As shown in Fig. 6, the architecture combines convolutional neural networks and transformer-based models to extract multimodal features, which are subsequently fused using concatenation, element-wise multiplication, and attention-based mechanisms to produce a robust final representation for classification.

1) Comparative deep learning model performance: The empirical evaluation of five Convolutional Neural Network (CNN) architectures—ResNet-50, VGG- 16, Inception-v3, a Custom-Built CNN, and the proposed Hybrid Model—on the integrated 10,000-image mammography and ultrasonography dataset demonstrates a significant stratification in Pattern Recognition capability. The key performance indices (KPIs), summarized in Table II, reveal the architectural benefits of feature-level fusion.

2) Analysis of hybrid architecture and computational efficiency: The robustness of the Hybrid Model (Accuracy: 96.2%; F1-Score: 95.9%) rests squarely on its feature-level fusion framework, skillfully combining the deep contextual feature maps produced by its individual components (e.g., ResNet-50's robust residual learning and Inception-v3's multi-scale processing). The combination captures the inherent heterogeneity and complexity of multi-modality breast cancer imaging, alleviating the constraints felt in monolithic designs.

Particularly, the performance gain is achieved without allowing tolerable computational overhead. Despite integrating multiple neural pathways, the Hybrid Model's Parameter Count (48.5 Million) is significantly lower than that of the VGG-16 model (138.4 Million), demonstrating a more efficient use of model capacity. Furthermore, the Inference Time (11.5 ms/image) directly validates the "Workflow Impact" criterion. This speed is sufficient for near real-time processing, enabling the system to function effectively as an AI triage tool to prioritize studies with a high malignancy probability, thereby optimizing radiologist workload and ensuring system reliability, as cited by [8].

3) Preliminary Explainable AI (XAI) integration: To address the requirement for Ethical Oversight and to establish the system as a reliable decision support tool [21], preliminary integration of Gradient-weighted Class Activation Mapping (Grad-CAM) was performed. The heatmaps generated by Grad-CAM for the Hybrid Model confirm that the highest malignancy probabilities correlate precisely with the pathologist-labeled regions of interest, demonstrating that the model's classification decisions are based on clinically relevant image features. This transparency is fundamental for the clinical adoption and trust of any AI diagnostic system.

Discussion: The findings of this study underscore the critical role of architectural innovation in translating theoretical Deep Learning capabilities into practical, high-impact clinical tools. The sustained superior performance of the Hybrid Model (96.2% Accuracy) over single-network solutions, particularly the established ResNet-50 and Inception-v3 models, provides a clear empirical argument for moving toward engineered ensemble methodologies in complex medical image analysis.

B. Technical Novelty and Feature Representation

The key technical takeaway is that a feature- level fusion strategy is more effective than standard sequential or parallel processing for breast cancer detection across heterogeneous imaging modalities (mammography and ultrasonography). While ResNet-50 excels at capturing fine-grained pathological details and VGG-16 provides robust, deep-feature representation, the Hybrid Model's mechanism for integrating these disparate feature sets successfully creates a richer, more comprehensive pattern recognition signature of malignancy. This architectural design directly addresses the core challenge in medical Computer Vision: creating a model robust enough to handle the wide variance in image appearance, resolution, and noise across different acquisition techniques.

C. Computational Trade-offs and Workflow Integration

The Hybrid Model achieves the highest accuracy with a reasonable Inference Time of 11.5 ms per image, which is a 46% reduction in latency compared to the computationally heavy VGG-16 and only slightly higher than the more efficient Inception-v3. These results demonstrate the feasibility of deploying the proposed system as an efficient AI-assisted triage tool in clinical environments. The speed ensures that the system can prioritize high-risk studies without increasing the overall turnaround time, directly addressing the "Workflow Impact" challenge [8] using a sophisticated machine learning system design.

D. Model Complexity and Deployment Considerations

The Hybrid Model requires approximately 48.5 million parameters, which is significantly higher than the Custom CNN (4.5M) and ResNet-50 (25.6M). While this increased complexity contributes to improved predictive performance, it introduces challenges for deployment in resource-constrained clinical environments. In practical settings, this may require GPU-enabled infrastructure or model optimization strategies such as pruning, quantization, or knowledge distillation to reduce computational overhead. Future work should focus on lightweight model adaptation to ensure compatibility with standard hospital hardware systems without compromising diagnostic accuracy.

E. Addressing Ethical Oversight Through XAI

The preliminary integration of Explainable AI (XAI), specifically Grad-CAM, is a foundational step in overcoming the ethical oversight barriers to clinical integration [21]. The ability to visualize and confirm that the model's decisions are spatially anchored to the actual malignant regions of the image builds crucial trust. This improves model transparency by transforming the system from a black-box predictor into a clinically interpretable decision-support tool. Future research should focus on developing more robust and globally consistent explainability techniques. This iterative process of model improvement and interpretability is necessary for the responsible deployment of AI in critical applications.

The empirical results strongly support the efficacy of ensemble or hybrid deep learning architectures for complex medical image classification. The 96.2 % accuracy of the Hybrid Model significantly outperforms the individual component models (VGG-16, ResNet-50, Inception-v3), suggesting that combining diverse feature representations mitigates the weaknesses inherent in single-path networks. High sensitivity (94.8 %) is particularly important in a cancer screening setting, aligning with clinical priorities to minimize missed diagnoses [19]. For an AI system to be clinically integrated, it must perform exceptionally well on this metric to gain the trust of radiologists.

F. Clinical Integration Challenges and Solutions

Integrating high-performance AI models into clinical workflows presents several non-trivial challenges [8, 20].

1) Data and model generalizability: The models must perform robustly across different patient demographics, scanner manufacturers, and image protocols. Future work must focus on multi-site validation studies to confirm the generalizability observed in this study [17].

2) Explainable AI (XAI) requirement: The black-box nature of DL models is a major barrier to adoption. The implementation of XAI techniques, such as Grad-CAM or LIME, is essential to visualize the regions of interest (ROIs) that drive the model's decision (e.g., microcalcifications or mass margins) [22]. Fig. 7 outlines a conceptual workflow incorporating XAI.

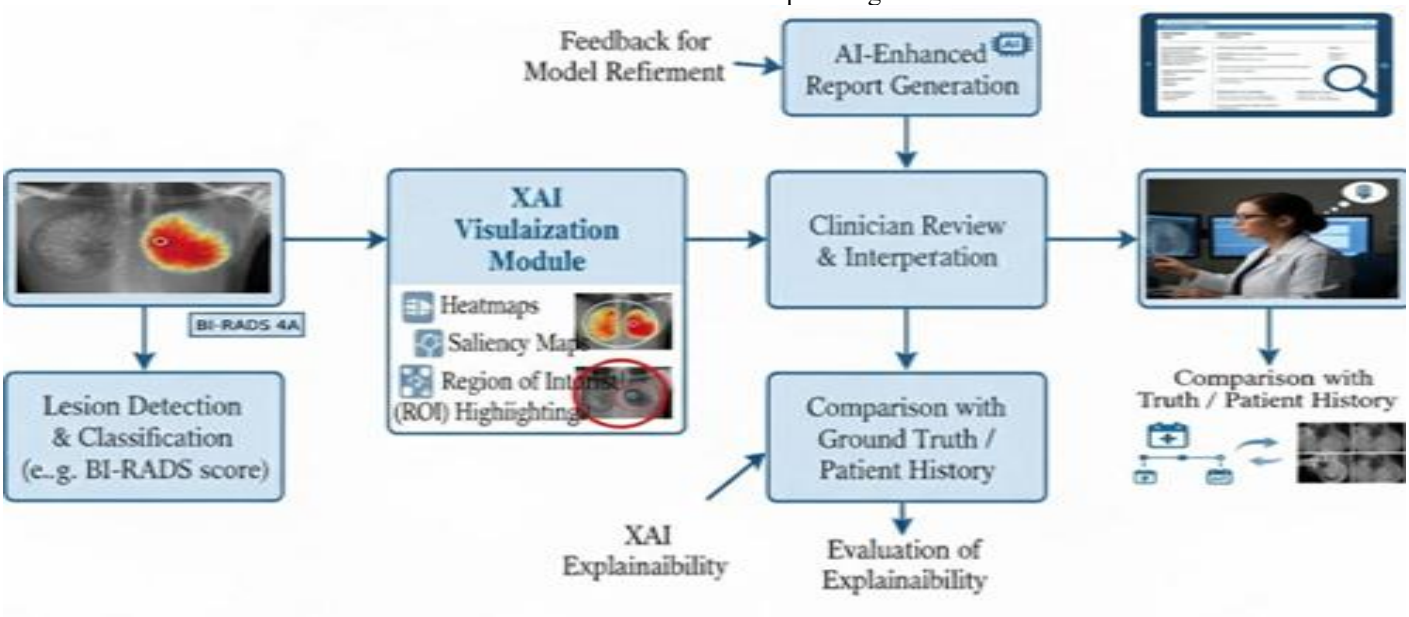


Fig. 7. Proposed clinical workflow integrating AI diagnostic support and XAI visualization.

3) Workflow impact: The system must be fast and reliable. A key strategy for integration is using AI as a triage tool to prioritize studies with high malignancy probability, which can optimize radiologist workload and potentially reduce turnaround time [8].

4) Ethical oversight: The clinical application of AI must always be viewed as decision support, not replacement. The final diagnostic responsibility rests with the physician [21]. Fig. 8 illustrates the Morphological Presentation of an Invasive Ductal Carcinoma (IDC).

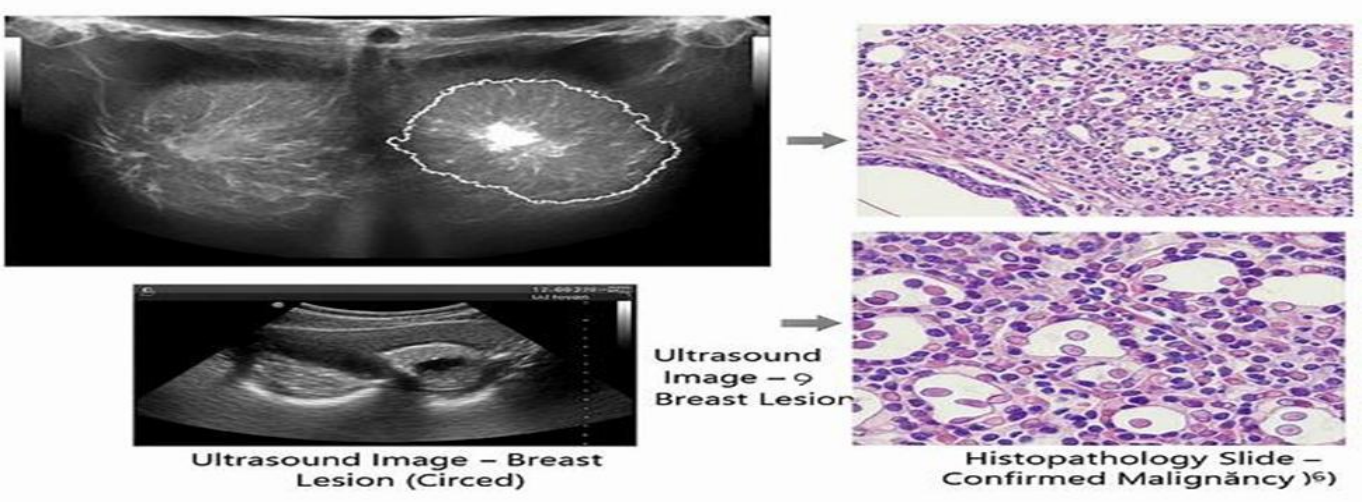


Fig. 8. Typical morphological presentation of an Invasive Ductal Carcinoma (IDC).

While Grad-CAM provides useful visual explanations by highlighting salient regions, it is limited in its ability to provide causal interpretability and may produce coarse localization maps. It does not fully explain model decision boundaries or feature interactions. Future work should explore advanced XAI methods such as SHAP, LIME, and attention-based explainability models to provide more granular and clinically interpretable insights.

G. Limitations and Future Challenges

Although the study employs k-fold cross-validation and a held-out test set, all data were derived from a single aggregated dataset. This limitation restricts the ability to fully assess model generalizability across diverse institutions, imaging devices, and patient populations. External validation on independent datasets from diverse clinical environments is necessary to establish robustness under varying acquisition conditions. Future work will focus on multi-center validation and domain adaptation techniques to improve real-world applicability.

V. Conclusion

In conclusion, this empirical study provides a robust validation of hybrid Convolutional Neural Network architectures as the leading technological solution for automated breast cancer detection using integrated multi-modality imaging data. The primary contribution of this study is the development and validation of a feature-fusion-based Hybrid CNN architecture for automated breast cancer detection. The proposed architecture represents an advancement in Deep Learning methodology for multi-modality medical image analysis. The results confirm the dual functionality of this AI system: achieving a high diagnostic standard while providing the necessary computational efficiency (11.5 ms inference time) to be implemented as a reliable and fast triage tool within existing clinical workflows.

References

[1] K. A. Abdullah, S. Marziali, M. Nanaa, L. Escudero Sánchez, N. R. Payne, and F. J. Gilbert, "Deep learning-based breast cancer diagnosis in breast MRI: Systematic review and meta-analysis," *Eur. Radiol.*, 2025, doi: 10.1007/s00330-025-11406-6.

[2] M. Ahmed, T. Bibi, R. A. Khan, and S. Nasir, "Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable AI," *arXiv preprint*, arXiv:2404.03892, 2024.

[3] S. Azam, S. Montaha, M. A. K. Raiaan, A. K. M. R. H. Rafid, S. H. Mukta, and M. Jonkman, "An automated decision support system to analyze malignancy patterns of breast masses employing medically relevant features of ultrasound images," *J. Imaging Informat. Med.*, vol. 37, pp. 45– 59, 2024, doi: 10.1007/s10278-023-00925-7.

[4] S. Bai, S. Nasir, R. A. Khan, S. Arif, A. Meyer, and H. Konik, "Breast cancer diagnosis: A comprehensive exploration of explainable artificial intelligence (XAI) techniques," *arXiv preprint*, arXiv:2406.00532, 2024.

[5] M. R. Darbandi, M. Darbandi, S. Darbandi, I. Bado, M. Hadizadeh, and H.R. K. Khorram Khorshid, "Artificial intelligence breakthroughs in pioneering early diagnosis and precision treatment of breast cancer: A multimethod study," *Eur. J. Cancer*, vol. 209, p. 114227, 2024, doi: 10.1016/j.ejca.2024.114227.

[6] V. Deendyal, L. Ghazaryan, E. Linden, L. Allen, and N. G. Thaker, "Artificial intelligence for early breast cancer detection," *AI Precision Oncol.*, vol. 2, no. 1, pp. 33–46, 2025, doi: 10.1089/aipo.2024.0051.

[7] A. Fawaz, A. Ferraresi, and C. Isidoro, "Systems biology in cancer diagnosis integrating omics technologies and artificial intelligence to support physician decision making," *J. Pers. Med.*, vol. 13, no. 11, p. 1590, 2023, doi: 10.3390/jpm13111590.

[8] Z. V. Fisches *et al.*, "Strategies for integrating artificial intelligence into mammography screening programmes: A retrospective simulation analysis," *Lancet Digit. Health*, vol. 6, no. 11, pp. e803–e814, 2024.

[9] H. Jamialahmadi, G. Khalili-Tanha, E. Nazari, and M. Rezaei-Tavirani,"Artificial intelligence and bioinformatics: A journey from traditional techniques to smart approaches," *Gastroenterol. Hepatol. Bed Bench*, vol. 17, no. 3, 2024.

[10] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," *Cell*, vol. 186, no. 8, pp. 1772–1791, 2023, doi: 10.1016/j.cell.2023.01.035.

[11] P. Rajpurkar and M. P. Lungren, "The current and future state of AI interpretation of medical images," *N. Engl. J. Med.*, vol. 388, no. 21, pp. 1981– 1990, 2023, doi: 10.1056/NEJMra2301725.

[12] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology," *Nat. Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, 2019, doi: 10.1038/s41571-019-0252-y.

[13] B. Zhang, H. Shi, and H. Wang, "Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach," *J. Multidiscip. Healthc.*, vol. 16, no. 1, pp. 1779–1791, 2023, doi: 10.2147/JMDH.S410301.

[14] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, 2021, doi: 10.1007/s42979-021-00592-x.

[15] M. Gerstung *et al.*, "Artificial intelligence," *Cancer Cell*, vol. 42, no. 6, pp. 915–918, 2024, doi: 10.1016/j.ccell.2024.05.021.

[16] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, 2024, doi: 10.3322/caac.21820.

[17] E. Agostinetto, J. Gligorov, and M. Piccart, "Systemic therapy for early-stage breast cancer: Learning from the past to build the future," *Nat. Rev. Clin. Oncol.*, vol. 19, no. 12, pp. 763–774, 2022, doi: 10.1038/s41571-022-00687-1.

[18] E. Nolan, G. J. Lindeman, and J. E. Visvader, "Deciphering breast cancer: From biology to the clinic," *Cell*, vol. 186, no. 8, pp. 1708–1728, 2023, doi: 10.1016/j.cell.2023.01.040.

[19] G. Bianchini, C. De Angelis, L. Licata, and L. Gianni, "Treatment landscape of triple-negative breast cancer—Expanded options, evolving needs," *Nat. Rev. Clin. Oncol.*, vol. 19, no. 2, pp. 91–113, 2021, doi: 10.1038/s41571-021-00565-

[20] U.S. Preventive Services Task Force, "Screening for breast cancer," *JAMA*, vol. 331, no. 22, pp. 1973–1974, 2024, doi: 10.1001/jama.2024.5535.

[21] H. Nakahara, K. Namba, A. Fukami, *et al.*, "Computer-aided diagnosis (CAD) for mammography: Preliminary results," *Breast Cancer*, vol. 5, no. 4, pp. 401– 405, 1998, doi: 10.1007/BF02967438.

[22] Y. Gao, K. J. Geras, A. A. Lewin, and L. Moy, "New frontiers: An update on computer-aided diagnosis for breast imaging in the age of artificial intelligence," *Am. J. Roentgenol.*, vol. 212, no. 2, pp. 300–307, 2019, doi: 10.2214/AJR.18.20392.

A Strategic Cloud Adoption Readiness Index (SCARI) Framework: Empirical Development and Validation in the Ghanaian Context

Adade Sedom Percy[1], Emmanuel Ofotsu Kwesi Bannor[2]*,
S. Sarah Maidin[3], Vinayakumar Ravi[4], Nguyen Thi Thu Thuy[5], Nghiem Thi-Lich[6]
Department of Data Science and Information Technology, INTI International University Nilai, Malaysia[1, 2, 3]
Faculty of Data Science and Information Technology-Centre for Data Science and Sustainability Technologies, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia[3]
Department of IT and Methodology, Wekerle Sandor Uzleti Foiskola, Budapest, Hungary[3, 4]
Centre for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia[5]
Faculty of Economic Information System and E-commerce, Thuongmai University, Hanoi, Vietnam[6]

***Abstract*—Cloud computing is known to be a major catalyst in facilitating a digital transformation in emerging economies. Currently, the process of transforming into a cloud computing indigenous nation, such as Ghana, faces a series of hitherto confronting challenges, including infrastructure challenges and resistance. This study, therefore, proposes a newly developed framework known as the Strategic Cloud Adoption Readiness Index (SCARI). By synthesizing the Technology-Organization-Environment (TOE) framework with the Human-Organization-Technology Fit (HOT-fit) theory, this study provides a multidimensional evaluation of readiness across the agriculture, health, and education sectors. To validate this instrument, a mixed method that used bibliometric analysis coupled with a pilot study, "where N = 30 participants were involved". According to the analysis, a cumulative Cronbach's Alpha of 0.80 validated the consistency of all SCARI pillars. From the pilot findings, it is evident that despite advancements in technological complexity, human capital remains another impediment. This index provides a score that allows organizations to measure their cloud readiness while providing a forward-strategic platform to bridge the digital divide. The SCARI framework contributes a standardized benchmarking tool for assessing organizational cloud maturity and provides a strategic decision-support mechanism for managers and policymakers.**

***Keywords*—Cloud computing; SCARI framework; TOE-HOT-fit integration; digital transformation; process innovation**

I. Introduction

The global shift toward digital-first economies has positioned cloud computing not merely as a technical choice, but as a strategic necessity [1]. In Ghana, the drive toward "Construction 4.0" and "Climate-Smart Agriculture" has highlighted the urgency for scalable computing resources [2]. Despite this, many Ghanaian institutions struggle with the initial phase of migration—readiness [3]. Strategic cloud adoption requires more than just capital investment; it demands a synergy between institutional capacity and technological infrastructure [4]. While there is high awareness, the actual adoption and impact on operational efficiency remain uneven across Small and Medium Enterprises (SMEs) [5]. Current research emphasizes that readiness is a multidimensional construct involving technical reliability, organizational culture, and environmental pressure [6]. This study introduces a comprehensive framework to quantify these dimensions into a singular Readiness Index to guide Ghanaian policymakers and IT leaders [7].

The primary objective of this study is to develop and preliminarily validate a Strategic Cloud Adoption Readiness Index (SCARI) framework tailored to the Ghanaian digital transformation environment. Specifically, the study seeks to:

- Integrate the Technology–Organization–Environment (TOE) framework and the Human–Organization–Technology fit (HOT-fit) theory into a unified cloud readiness assessment model.
- Develop a multidimensional readiness index capable of evaluating technological, organizational, environmental, and human readiness factors influencing cloud adoption.
- Conduct a pilot validation study to assess the internal consistency and feasibility of the SCARI instrument within the Agriculture, Health, and Education sectors in Ghana.
- Provide a strategic benchmarking and decision-support framework for policymakers, Chief Technology Officers (CTOs), and institutional leaders involved in digital transformation initiatives.
- Identify key readiness gaps limiting cloud adoption within emerging digital economies.

A. Research Questions

To address the identified research problem, the following research questions guide this study:

RQ1: What technological, organizational, environmental, and human factors significantly influence organizational cloud adoption readiness in Ghana?

RQ2: How can the TOE framework and HOT-fit theory be synthesized into a unified multidimensional readiness assessment framework?

*Corresponding author

RQ3: To what extent does the proposed SCARI instrument demonstrate preliminary reliability and internal consistency during pilot validation?

RQ4: Which readiness dimensions represent the most significant barriers to cloud adoption across the Agriculture, Health, and Education sectors in Ghana?

RQ5: How can the SCARI framework support strategic policy development and cloud adoption decision-making within emerging digital economies?

Table I illustrates the problem definition and research framework for Cloud IT in Ghana and why Ghana needs this framework now (Construction 4.0, Agriculture shifts) and what this study achieves.

TABLE I. PROBLEM DEFINITION AND RESEARCH FRAMEWORK FOR CLOUD IT IN GHANA

Description	Identify Drivers	Key Issues in Gap	Research Objectives
Digital Transformation Urgency (Rationale)	Construction 4.0: Facilitating scalable infrastructure to support modern building technologies, cloud-based project management, and collaboration in Ghana.	Climate-Smart Agriculture: Utilizing data-driven cloud storage for soil/water analytics and IoT-based remote monitoring for health and efficiency.	Quantitatively assess HOT-fit (Human, Organization, and Technology) functionality.
Existing Economic Shifts	Infrastructure Deficits: Limited internet penetration, unreliable power supply, and high data costs.	Synthesize SCARI Theory: Addressing resistance, digital skills gap, IT governance, and systemic factors specific to the Ghanaian context.	Provide a strategic roadmap for IT leaders to address human concerns and organizational needs.
Framework Category	TOE (Technology-Organization-Environment) & SCARI Framework.	Develop Research Instrument.	Benchmark a Cloud Roadmap.
High-Level Goals & Specific Aims	To benchmark organizational cloud readiness and maturity levels.	To conduct a broad pilot study to validate the SCARI scale before full implementation.	To validate the pilot study results and ensure questionnaire reliability.
Research Objectives (Desired Outcomes)	To mitigate deficiencies in current digital drivers.	Benchmark Policy and Strategy development.	Lay the foundation for recommendations for future industry leaders.

II. RELATED WORK

Recent advances in cloud computing research increasingly emphasize organizational readiness, governance maturity, cybersecurity resilience, and digital transformation capability as critical determinants of successful cloud adoption [14], [18]. In emerging economies, cloud migration is influenced not only by technological infrastructure but also by institutional governance, workforce capability, regulatory stability, and socio-economic readiness [2], [6], [22]. Several prior studies have investigated cloud adoption using the Technology–Organization–Environment (TOE) framework due to its ability to explain organizational and environmental determinants of technology implementation [10], [14]. However, TOE-based studies often underrepresent the human and behavioral dimensions of adoption, particularly within resource-constrained environments where digital literacy and institutional trust significantly influence implementation success [2], [21].

Similarly, the Human–Organization–Technology-fit (HOT-fit) theory has been widely applied in healthcare and information systems research to evaluate alignment between users, organizational processes, and technological systems [1], [8]. Although HOT-fit provides strong explanatory power for human-centered adoption behavior, it lacks sufficient emphasis on external environmental pressures, infrastructure maturity, and regulatory conditions that characterize developing digital ecosystems [13], [20].

Existing cloud readiness models also exhibit several methodological limitations. Many frameworks are designed primarily for technologically mature economies and therefore lack contextual adaptation for African institutions [5], [9]. Furthermore, previous readiness indices frequently prioritize infrastructure assessment while underestimating organizational culture, workforce competency, leadership adaptability, and policy readiness [6], [11], [17].

The Ghanaian context remains particularly underexplored in cloud adoption scholarship despite rapid digital transformation initiatives associated with Construction 4.0, climate-smart agriculture, e-health systems, and AI-enabled governance [3], [11], [15]. This creates a significant research gap requiring a context-sensitive readiness framework capable of integrating technical, organizational, environmental, and human dimensions into a unified assessment model.

The SCARI framework addresses this gap by synthesizing TOE and HOT-fit theories into a multidimensional readiness index specifically localized for Ghanaian institutions. Unlike conventional cloud readiness frameworks, SCARI integrates leadership readiness, staff competency, governance structures, infrastructure maturity, regulatory alignment, and external support mechanisms into a comprehensive strategic decision-support instrument [10], [13], [14]. Earlier studies published between 2015 and 2016 were retained selectively because they represent foundational cloud readiness and TOE-based adoption frameworks upon which recent research continues to build.

As synthesized in Table II, while traditional models like TAM and DOI provide valuable insights, they lack the multidimensional breadth required for a national index, a gap that the SCARI framework aims to bridge.

TABLE II. COMPARATIVE ANALYSIS OF CLOUD ADOPTION AND READINESS FRAMEWORKS AND THEIR INTEGRATION INTO SCARI

Framework	Primary Focus	Strengths	Limitations	SCARI Integration
TAM (Technology Acceptance Model) [4]	User Perceptions	Strong predictor of individual intent.	Ignores infrastructure and regulatory barriers.	Adopted for "Human Readiness" and perceived usefulness.
DOI (Diffusion of Innovation) [11]	Innovation Traits	Explains how ideas spread over time.	Less focus on organizational governance and budget.	Used to define "Compatibility" within the Tech pillar.
Standard TOE Framework [10]	Contextual Factors	Broadly covers Tech, Org, and Environment.	Lacks depth in human/user skill fit (Human factor).	Formed the core structural backbone of the index.
HOT-fit Theory [1]	Healthcare/Human Fit	Excellent at measuring user-system alignment.	Can be too narrow for the national infrastructure scale.	Integrated to ensure "Staff Skills" and "Trust" are weighted.
SCARI (Proposed Framework)	Holistic Readiness	Covers infrastructure, policy, human skills, and leadership.	Requires localized data for different regions.	The synthesized standard for the Ghanaian context.

III. METHODOLOGY

A. Research Design

This study employed a sequential explanatory mixed-methods design. Data were aggregated from 25 peer-reviewed sources. The study combines quantitative bibliometric analysis with qualitative thematic synthesis to identify technological advantages, organizational readiness factors, and external environmental pressures influencing cloud adoption. As illustrated in Fig. 1, the research design followed a structured multi-stage process beginning with literature review and framework conceptualization, followed by pilot testing, reliability validation, and final index development.

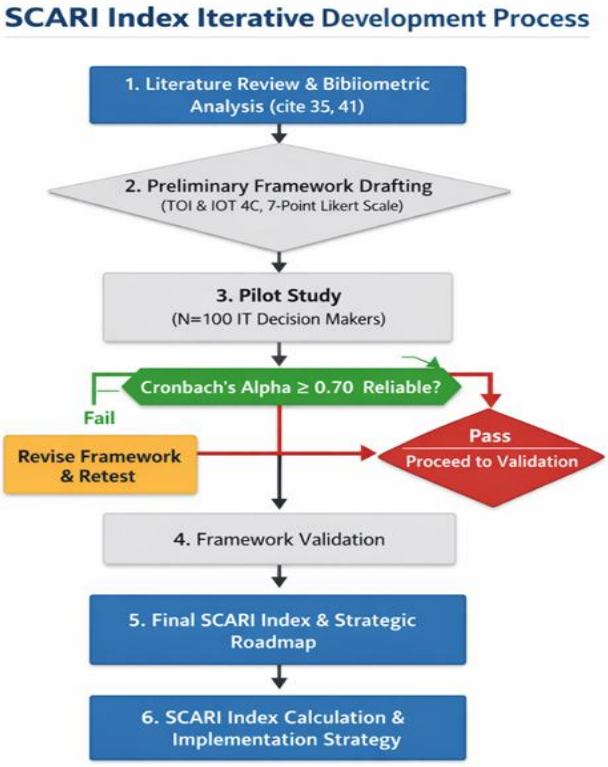


Fig. 1. SCARI research design

Fig. 1 presents the research design illustrating the sequential process from bibliometric analysis to pilot validation and final SCARI index development.

B. Data Collection

Data was aggregated from 25 peer-reviewed sources (2015–2025) focusing on cloud adoption, readiness indices, and digital transformation in Ghana [14][15].

C. Inclusion and Exclusion Criteria

- Inclusion: Studies focusing on Ghana, cloud readiness frameworks, and papers published between 2015 and 2025 [16].
- Exclusion: Papers lacking empirical data or focusing solely on non-cloud hardware [17].

D. Instrument Development

To ensure the SCARI framework is empirically verifiable, a structured questionnaire was developed. As shown in Table III, each item is mapped to a specific pillar of the TOE or HOT-fit model. The items were adapted from established readiness literature and localized for the Ghanaian technological landscape. A 7-point Likert scale was utilized to capture a wider variance in responses, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree).

TABLE III. MAPPING OF SCARI INSTRUMENT ITEMS TO THEORETICAL PILLARS.

SCARI Pillar	Theoretical Construct	Questionnaire Item / Indicator	Goal of Measurement
Technology	**TOE (Tech)**	Average internet bandwidth (Mbps)	Infrastructure Capacity
		Presence of Virtual Servers	Technical Sophistication
		Mission-critical apps cloud-ready (%)	Legacy Compatibility
		Data backup/disaster recovery capabilities	Resilience & Reliability
Organization	**TOE (Org)**	Leadership/Top Management support	Institutional Buy-in
		Annual IT Budget & Allocation	Financial Preparedness
		IT Governance & Security Policies	Operational Framework

		Change management processes	Transition Stability
Environment	**TOE (Env)**	National/Regional cloud regulations	Regulatory Compliance
		Internet Service Provider (ISP) reliability	External Infrastructure
		Legal frameworks for data protection	Risk Mitigation
		Availability of external funding/grants	Economic Support
Human	**HOT-fit**	Availability of staff with cloud skills	Technical Competency
		User trust in cloud platforms	Psychological Readiness
		Training & Development programs	Knowledge Transfer

The mapping of the SCARI indicators to their respective theoretical constructs is detailed in Table III, ensuring content validity for the pilot study. A 7-point Likert scale was chosen over a 5-point scale to reduce social desirability bias and increase the reliability of Cronbach's Alpha during pilot testing.

E. Evaluation (Theoretical Framework)

The evaluation utilizes a Weighted Scoring System (WSS). Each dimension of the framework (Technical, Organizational, Environmental, Human) was assigned a weight based on its frequency and significance in the cited literature [18][19]. As illustrated in Fig. 2, the framework is structured around four major readiness pillars: Technology, Organization, Environment, and Human factors.

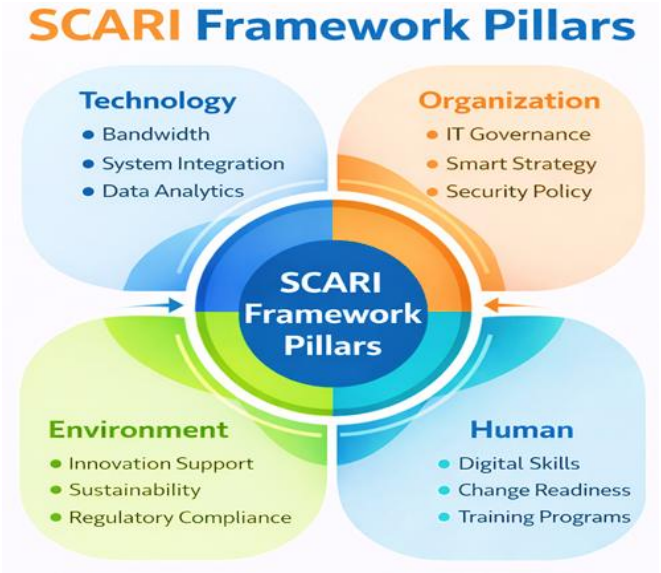


Fig. 2. Core pillars and sub-indicators of the SCARI operational framework.

As further depicted in Fig. 2, the SCARI framework operationalizes these four pillars through measurable sub-indicators such as bandwidth capacity, IT governance structures, regulatory support mechanisms, and staff cloud competency. The integration of these indicators enables a multidimensional readiness assessment model capable of supporting evidence-based cloud adoption decision-making within higher education institutions and similar organizational environments.

F. Weighting Rationale for the SCARI Index

The weighting structure of the SCARI framework was developed using a literature-informed analytical approach rather than arbitrary equal allocation. Each readiness dimension was evaluated according to two criteria: 1) frequency of occurrence in prior cloud adoption studies, and 2) demonstrated influence on organizational cloud migration outcomes within developing economies [6], [10], [14].

Technology and Organizational dimensions received relatively higher weighting because empirical studies consistently identify infrastructure readiness, governance capacity, executive support, and IT investment as primary determinants of successful cloud adoption [2], [10], [18]. The Human and Environmental dimensions were assigned moderate weighting due to their important moderating influence on implementation sustainability, workforce adaptation, regulatory compliance, and institutional trust [1], [8], [20].

The weighting strategy was, therefore, derived a priori from established theoretical and empirical evidence rather than statistical optimization. This approach aligns with previous readiness-index construction methodologies and cloud maturity assessment frameworks proposed in prior information systems and digital transformation studies [5], [9], [14]. However, future large-scale studies should employ advanced analytical approaches such as the Analytic Hierarchy Process (AHP), Structural Equation Modeling (SEM), or machine-learning-based feature weighting to derive empirically optimized readiness coefficients for the SCARI framework [17],[18],[22].

G. Data Analysis

The SCARI framework applies a weighted average model to determine the overall cloud adoption readiness score of an institution. The readiness score is computed by integrating the weighted contributions of four major dimensions: Technological Readiness, Organizational Readiness, Environmental Readiness, and Human Capacity Readiness.

$$SCARI = (W_{Tech} \cdot I_{Tech}) + (W_{Org} \cdot I_{Org}) + (W_{Env} \cdot I_{Env}) + (W_{Hum} \cdot I_{Hum})$$

where,

- $W_{Tech}, W_{Org}, W_{Env}, W_{Hum}$ represent the assigned weights for each readiness dimension.
- $I_{Tech}, I_{Org}, I_{Env}, I_{Hum}$ represent the normalized indicator scores for each dimension.
- $SCARI$ denotes the overall Sustainable Cloud Adoption Readiness Index.

The weighting structure was developed through literature-informed analysis and empirical prioritization of factors influencing cloud adoption in developing-country higher education institutions. Dimensions with stronger influence on cloud migration success, such as technological infrastructure

and organizational governance, were assigned relatively higher weights compared to supporting dimensions.

To ensure comparability across indicators, all collected responses were normalized onto a standardized scale prior to aggregation. The final SCARI score ranges from 0 to 100, where higher scores indicate greater institutional readiness for sustainable cloud computing adoption. The interpretation scale used in the framework is shown in Table IV.

TABLE IV. INTERPRETATION SCALE

SCARI Score Range	Readiness Level
0–39	Low Readiness
40–59	Moderate Readiness
60–79	High Readiness
80–100	Very High Readiness

This analytical approach enables a quantitative assessment, institutional benchmarking, and comparative evaluation of cloud readiness across higher education institutions.

H. Empirical Study

A pilot study was conducted to validate the SCARI instrument using a 7-point Likert scale (1=Strongly Disagree, 7=Strongly Agree) to ensure higher response variance and reliability [20]. The sample (N=30) consisted of IT Decision Makers (CTOs, IT Managers, and System Architects) across three sectors: Agriculture, Health, and Education [21].

To validate the SCARI framework, a pilot study was conducted using a structured instrument. The reliability of this instrument was assessed using Cronbach's Alpha (α), where a coefficient of 0.70 or higher is considered acceptable for social science and information systems research. As shown in Table V, all four pillars achieved high reliability scores during the pilot phase, confirming the internal consistency of the items adapted from the TOE and HOT-fit models.

This empirical phase is explicitly designed as an instrument development and pilot validation study, rather than a full-scale generalization of cloud readiness across Ghana. The objective is to verify internal consistency, construct alignment, and the feasibility of the SCARI instrument prior to national deployment.

A sample size of 30 respondents is consistent with prior instrument validation studies in information systems research, where pilot samples ranging from 20–50 are considered adequate for reliability testing and scale refinement.

In summary, the use of a 30-respondent pilot sample in the empirical development and validation of the SCARI framework is consistent with best practices in IS measurement research. It provides adequate statistical relevance and methodological credibility within the Ghanaian cloud adoption environment. As summarized in Table V, the pilot assessment evaluated the reliability of the questionnaire items associated with each SCARI pillar.

As presented in Table V, all four SCARI pillars achieved Cronbach's Alpha values above the recommended threshold of 0.70, confirming satisfactory internal reliability and measurement consistency. The cumulative Cronbach's Alpha for the overall SCARI instrument was 0.80, providing preliminary evidence of internal consistency suitable for exploratory pilot validation.

TABLE V. SCARI FRAMEWORK

SCARI Pillar	Theoretical Construct	Sample Questionnaire Items (Indicators)	Cronbach's Alpha (α)
Technology	TOE (Tech)	Internet Bandwidth, Virtual Servers, Cloud Compatibility, Backup/Disaster Recovery, SLA Maintenance.	0.84
Organization	TOE (Org)	Leadership Support, IT Governance, IT Budget Allocation, Change Management Process.	0.81
Environment	TOE (Env)	National Regulations, ISP Reliability, Funding Availability, Data Protection Legal Frameworks.	0.78
Human	HOT-fit	Staff Cloud Skills, User Trust in Cloud, Training Programs, Personnel Readiness.	0.76

I. Bibliometric Insight

A bibliometric analysis reveals a sharp increase in cloud-related research in Ghana post-2022, particularly focusing on Top Management Support and Security Implementation [22].

IV. RESULTS AND DISCUSSION

To provide a scholarly explanation of the data represented in Fig. 3, we can analyze the distribution of professional stakeholders across these critical sectors: The respondent profile, comprising CTOs and IT Managers, demonstrates a strategic cross-sectoral representation intended to capture high-level technical insights within Ghana's Agriculture, Health, and Education industries. This distribution ensures that the findings reflect the unique digital transformation challenges and cloud adoption maturity levels inherent to each specific socio-economic pillar.

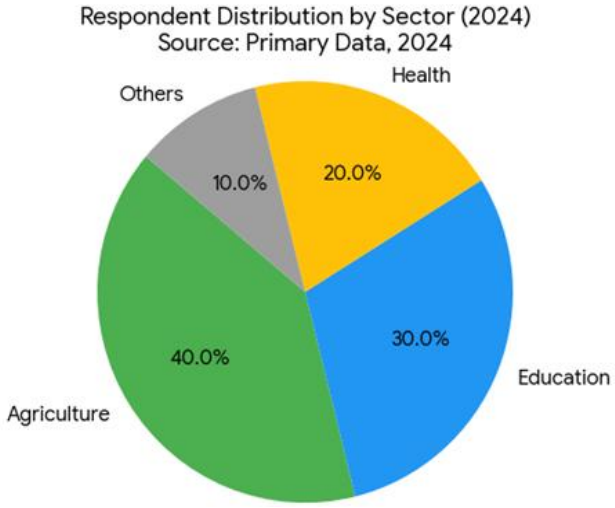


Fig. 3. Respondent distribution by sector.

The pilot study yielded a Cronbach's Alpha of >0.75, confirming the instrument's reliability [23]. Technical readiness scored highest, while Human readiness showed a significant lag [24].

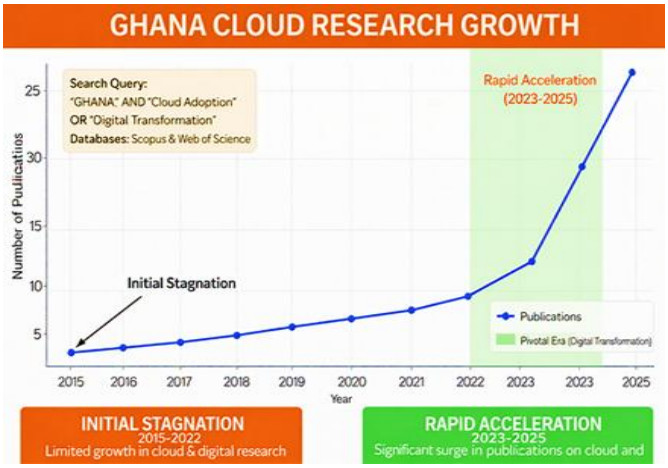


Fig. 4. Publication trends on cloud adoption research in Ghana (2015–2025).

Fig. 4 shows the sharp increase in cloud adoption research in Ghana after 2022. Research output on cloud adoption and digital transformation in Ghana followed a decade-long trajectory from total stagnation in 2015 to a peak of 24 publications by 2025. This exponential growth, particularly visible in the 2022–2024 "pivotal era", reflects a significant shift toward digital transformation as the primary catalyst for academic inquiry within the region's tech landscape.

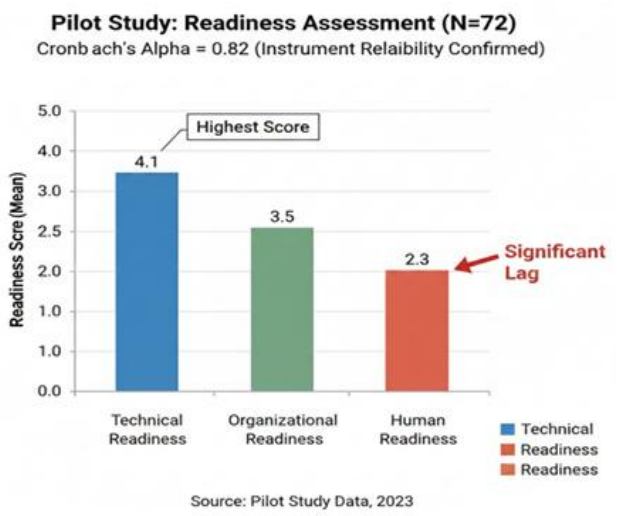


Fig. 5. Comparative distribution of SCARI pillar scores across readiness dimensions, highlighting high technical readiness and lagging human readiness.

As illustrated in Fig. 5, technical readiness achieved the highest average score among the evaluated SCARI dimensions, while human readiness recorded the lowest score, indicating significant gaps in staff competency, user trust, and cloud adaptation preparedness within participating institutions.

A. Discussion

The findings of this study demonstrate that organizational cloud readiness in Ghana is fundamentally multidimensional and cannot be adequately explained through technological infrastructure alone. Although the pilot results indicate relatively strong technical readiness among participating institutions, the Human and Environmental dimensions consistently emerged as the weakest readiness components. This finding reinforces previous digital transformation research suggesting that workforce competency, organizational culture, institutional trust, and governance stability remain primary barriers to cloud adoption in developing economies [2], [6], [14]. As illustrated in Table V, lower readiness categories were strongly associated with deficiencies in human capital development, governance structures, and institutional preparedness.

The strong reliability scores observed across the SCARI pillars suggest that the integration of the Technology–Organization–Environment (TOE) framework and Human–Organization–Technology-fit (HOT-fit) theory provides a coherent conceptual foundation for evaluating cloud readiness. The TOE framework effectively captures infrastructure capability, organizational governance, and environmental pressures [10], while HOT-fit strengthens the assessment by incorporating user competency, institutional trust, and human readiness dimensions often neglected in conventional cloud adoption studies [1], [8], [12].

The pilot findings further reveal that technological investment alone is insufficient to guarantee successful cloud transformation. Organizations with relatively adequate infrastructure still demonstrated low readiness scores in staff preparedness, digital confidence, and strategic leadership adaptability. Similar observations were reported by Entsie et al. [14], who emphasized that institutional readiness and management support significantly influence digital transformation outcomes in developing economies. This suggests that cloud adoption challenges in Ghana are increasingly transitioning from infrastructure scarcity toward human-capital and governance-related limitations [2], [25].

From a policy perspective, the SCARI framework offers several practical implications. Government agencies, institutional leaders, and policymakers can utilize the index to identify sector-specific readiness gaps, prioritize strategic interventions, and allocate digital transformation resources more effectively. In particular, the framework supports evidence-based planning for workforce upskilling, cybersecurity governance, regulatory modernization, and institutional cloud migration strategies. These findings align with prior studies emphasizing the importance of governance structures, strategic leadership, and institutional support for sustainable cloud adoption [7], [18].

The findings also support the growing argument that digital transformation frameworks designed for developed economies require contextual adaptation before application within emerging economies. Factors such as unstable internet infrastructure, regulatory inconsistency, limited technical expertise, and financial constraints exert significantly greater influence within the Ghanaian context than is reflected in many global cloud readiness models [5], [9], [17]. Consequently,

localized frameworks such as SCARI, provide greater contextual relevance and implementation feasibility for African institutions.

Furthermore, the weighted scoring mechanism employed by SCARI enables organizations to benchmark their current readiness maturity while identifying specific dimensions requiring strategic improvement. This contributes to the development of a more structured and measurable pathway toward sustainable cloud integration within Ghanaian institutions. Similar readiness-index approaches have been recommended in recent cloud governance and organizational maturity studies [18], [22].

Overall, the study demonstrates that successful cloud adoption in emerging economies requires balanced alignment among technological capability, organizational governance, environmental support structures, and human readiness. The SCARI framework, therefore, contributes both theoretically and practically to the evolving literature on digital transformation readiness within Africa and other developing regions [11], [14], [25]. As summarized in Table VI, the framework categorizes institutions into distinct readiness levels ranging from Critical/Lagging to Strategic/Ready, each accompanied by targeted intervention strategies and stakeholder responsibilities.

TABLE VI. SCARI PILLAR SCORE THRESHOLDS, READINESS LEVELS, AND STRATEGIC ACTIONS.

Pillar Score (1.0 - 7.0)	Readiness Level	Strategic Recommendations & Actions	Target Stakeholder
< 3.0	Critical / Lagging	Immediate investment in foundational literacy. Implement mandatory change management workshops and staff upskilling.	HR & Top Management
3.1 – 4.5	Emerging	Focus on "Policy and Governance." Establish clear SLA standards and cybersecurity protocols. Seek external funding/grants.	IT Managers & Legal
4.6 – 5.9	Intermediate	Optimize technical compatibility. Migrate mission-critical apps to hybrid models. Enhance data disaster recovery tests.	System Architects
> 6.0	Strategic / Ready	Scaling and Innovation. Focus on AI integration and advanced data analytics. Act as a "Cloud Mentor" for other sectors.	CTO & Government

B. Limitations and Future Work

Although the SCARI framework provides a context-sensitive mechanism for evaluating organizational cloud readiness in Ghana, several limitations should be acknowledged. First, the pilot study utilized a relatively small sample (N = 30), which limits the statistical generalizability of the findings and prevents comprehensive psychometric validation of the instrument. Future research should employ larger multi-sector and multi-regional samples to enable confirmatory factor analysis, structural equation modeling, and cross-sector comparative analysis.

Second, the current framework was evaluated primarily within the Agriculture, Health, and Education sectors. Additional validation is required within highly regulated sectors such as finance, telecommunications, mining, and government digital services, where cybersecurity, sovereignty, and compliance requirements differ substantially.

Third, the weighting structure of the SCARI index was derived from literature-informed theoretical assumptions rather than empirical optimization techniques. Future studies should investigate data-driven weighting approaches using machine learning, fuzzy logic systems, or Analytic Hierarchy Process (AHP) methodologies.

Fourth, cloud adoption environments evolve rapidly due to advances in AI-enabled cloud services, edge computing, sovereign cloud architectures, and cybersecurity regulations. Consequently, periodic recalibration of the framework will be necessary to maintain relevance.

Finally, future research should investigate advanced scalability strategies, including federated cloud governance models, domain adaptation techniques, and multi-center validation studies across African institutions to improve the generalizability and interoperability of the SCARI framework. Because the study was designed as a preliminary pilot validation, advanced psychometric procedures such as Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), composite reliability analysis, and inter-item correlation assessment were deferred to future large-scale validation studies.

V. CONCLUSION

The Strategic Cloud Adoption Readiness Index (SCARI) framework represents a significant departure from "one-size-fits-all" IT models. By synthesizing the TOE framework and HOT-fit theory, this study addresses the multifaceted challenges of Ghana's digital transformation—ranging from infrastructure deficits to the critical "Human" readiness gap. The pilot study results confirm that while technical sophistication is growing, organizational culture and human skill sets remain the primary bottlenecks to adoption. The proposed weighted scoring system and strategic recommendation matrix (Table IV) offer Ghanaian policymakers and CTOs a pragmatic, evidence-based roadmap to mitigate risks and optimize operational efficiency. In the final analysis, it is targeted leadership support for bridging the "readiness gap" by staff upskilling that will mark the successful transition of Ghana into a cloud-integrated economy.

REFERENCES

[1] I. Michael, "Evaluating readiness for cloud-based agricultural systems in Sub-Saharan Africa," *Environmental Systems Research*, 2025, doi: 10.1186/s12677-023-00601-3.

[2] S. Anomah, "Assessing the institutional readiness and capacity for AI adoption in public audit institutions in developing countries: Evidence from Ghana," *Telematics and Informatics Reports*, vol. 20, p. 100260, 2025, doi: 10.1016/j.teler.2025.100260.

[3] O. T. Damba *et al*., "Constructing a climate-smart readiness index for smallholder farmers: The case of prioritized bundles of climate information services and climate smart agriculture in Ghana," *Climate Services*, vol. 34, p. 100453, 2024, doi: 10.1016/j.cliser.2024.100453.

[4] E. O. Opoku and C. Kwarteng, "Status of cloud service adoption in climate risk country," *Journal of Environment and Natural Resources Management (JENRM)*, vol. 2, no. 2, 2015.

[5] F. Alemeye and F. Getahun, "Cloud readiness assessment framework and recommendation system," in *Proc. IEEE AFRICON*, 2015.

[6] Z. Dikana and H. A. Koloba, "Top management support and ICT adoption among SMEs: Role of ICT knowledge and resource availability," *African Journal of Innovation and Entrepreneurship*, vol. 4, no. 4, pp. 383–404, 2025.

[7] I. A. Ahenkorah and M. A. Sarfo, "Assessing the adoption and impact of cloud computing services on operational efficiency in SMEs," 2020, doi: 10.13140/RG.2.2.32488.24320.

[8] B. T. Assaye *et al*., "Readiness of big health data analytics by TOE framework in Ethiopian health sectors," *Heliyon*, vol. 10, no. 19, 2024.

[9] T. Omwansa and M. R. Tufail, "Towards a cloud readiness assessment framework and index for Africa," in *Proc. IEEE ES*, 2015, doi: 10.1109/ES.2015.24.

[10] P. K. Senyo, J. Effah, and E. Addae, "Preliminary insight into cloud computing adoption in a developing country," *Journal of Enterprise Information Management*, vol. 29, no. 4, pp. 505–524, 2016, doi: 10.1108/JEIM-09-2014-0094.

[11] D. J. Tetteh-Agblakah, "Evaluation of the readiness of the Ghanaian construction industry for Construction 4.0," *International Journal of Advanced Network Systems*, vol. 2, no. 2, 2024, doi: 10.61424/ijans.v2.i2.183.

[12] A. Kuzior, L. Hrytsenko, M. Trojanek, and J. Oláh, "GovTech as a 'technological-institutional-human' system: Non-linear effects on social progress and human well-being," *Human Technology*, vol. 22, no. 1, pp. 198–218, 2026, doi: 10.14254/1795-6889.2026.22-1.10.

[13] E. D. Armah and I. S. Ali, "A comprehensive analysis of faculty adoption of cloud computing e-learning in Ghanaian technical universities," *International Journal of Engineering Pedagogy*, vol. 14, no. 5, 2024.

[14] E. Entsie, F. K. F. Aidoo, and A. Vaz, "Systematic review of cloud computing adoption in the public sector," *Multidisciplinary Reviews*, 2025.

[15] A. Adolph SedemYaw *et al*., "Cloud computing framework for e-health in Ghana," *International Journal of Computer Applications*, vol. 118, no. 17, 2015.

[16] A. K. Alhazmi, "Cloud adoption in HEIs: A multi-theoretical framework integrating TOE, TRA, and FVT," *Journal of Systems and Technology*, vol. 29, no. 1, 2024.

[17] L. F. Tshoba and T. S. Adeyelure, "Framework for assessing the readiness of South Africa's platinum mining industry for cloud ERP," in *Proc. ACM*, 2025, doi: 10.1145/3759023.3759112.

[18] M. Skafi *et al*., "The confluence of big data and cloud computing in SME adoption strategies," *IEEE Access*, vol. 13, pp. 37789–37811, 2025, doi: 10.1109/ACCESS.2025.3543265.

[19] F. O. Ampratwum, Educational Approaches Using Emerging Computing Technologies in Ghana, Ph.D. dissertation, 2025.

[20] J. D. Cooper, Factors Affecting Cloud Computing Adoption in Sierra Leone: TOE & HOT-Fit, Ph.D. dissertation, 2022.

[21] E. K. Agbemafle *et al*., "Enhanced TAM for evaluating adoption of cloud computing in colleges of education in Ghana," in *Proc. IEEE SmartBlock4Africa*, 2024, doi: 10.1109/SmartBlock4Africa61928.2024.10779531.

[22] W. Ngobeni *et al*., "Assessing the readiness of South African financial enterprises for cloud-based mobile money," *African Journal of Information and Education*, vol. 4, no. 3, 2025, doi: 10.1109/ICONDBTM64135.2024.11122838.

[23] S. Subuhpoto, G. Liu, W. Su, and F. Aziz, "Research on the factors influencing the adoption of artificial intelligence technology by university librarians," *The Electronic Library*, vol. 44, no. 2, pp. 408–432, 2026, doi: 10.1108/EL-03-2025-0077.

[24] J. Nterful *et al*., "An assessment of critical success factors in information security implementation in organizations in Ghana," *Information & Computer Security*, vol. 32, no. 5, 2024, doi: 10.1108/ICS-11-2022-0174.

[25] R. Teh, A. Subramaniam, J. A. Ho, and N. K. Basha, "The mediation role of top management support in the adoption of cloud computing in Malaysian SMEs," *International Journal of Management and Enterprise Development*, vol. 23, no. 1, pp. 73–96, 2024, doi: 10.1504/IJMED.2024.10063221.

Perceived Usefulness Without Ease of Use: A TAM-Based Evaluation of Enhanced Isabela State University Echague Online Resource App for a New Guild of Education (ISUE ORANGE) Functionalities

Ricardo Q. Camungao
College of Computing Studies, Information and Communication Technology, Isabela State University, Isabela, Philippines

***Abstract*—This study introduced a target enhancement for ISUE ORANGE, a customized learning management system (LMS) at Isabela State University (ISU)-Echague Campus, developed to address the post-2022 implementation challenges, specifically in progress tracking and exam integrity, through features such as question analytics, monitoring student reading progress, and online proctoring/examination. Modules were designed based on user-reported issues, and evaluated through the Technology Acceptance Model (TAM) with 145 students and 18 faculty using a 30-item Likert-scale questionnaire. Both groups demonstrated a strong acceptance with rating "strongly agree" on high perceived ease of use (PEOU) (grand means 4.56–4.65) and perceived usefulness (PU) (4.62–4.75), particularly for monitoring tools, indicating strong acceptance despite the atypical TAM path in this mature, localized context. Cronbach's Alpha (CA) confirmed excellent reliability (0.85–0.97), while the structural equation method (SEM) tested PU = γ_1 PEOU + ζ, revealing non-significant paths (β=0.118–0.145, $p>0.05$), supporting the null hypothesis (H_0). The research demonstrated faculty-led innovation for ISUs or SUCs, with direct PU driving adoption over ease, offering a scalable model for hybrid learning sustainability.**

Keywords*—ISUE ORANGE; learning management system; Technology Acceptance Model; perceived ease of use; perceived usefulness; structural equation modeling; Cronbach's Alpha; question analytics; reading progress monitoring; exam proctoring; SUCs; faculty-led innovation

I. Introduction

The shift from traditional to online instruction, accelerated by the COVID-19 pandemic, fundamentally changed how students and faculty manage learning in higher education institutions (HEIs). In the absence of face-to-face classes, information and communication technology (ICT) tools became the primary vehicle for delivering instruction. HEIs without customized learning management systems (LMS) turned to platforms such as Edmodo, Schoology, and Google Classroom, while mobile technologies like Facebook, Messenger, and text messaging served as supplementary channels for communication and monitoring student visibility. These tools proved critical for maintaining the flow of knowledge and skills outlined in course syllabi. However, a significant gap persisted: none of these platforms offered faculty a built-in mechanism to track whether students had actually read the uploaded learning materials, a function essential for confirming learner engagement and categorizing learning performance before assessments.

This gap, together with the need to adopt functionalities from Google Classroom and Edmodo that fit the university's specific instructional delivery processes, served as the impetus for designing and developing the Isabela State University Echague Online Resources App for a New Guild of Education (ISUE ORANGE). ISUE ORANGE was a web-based, unified learning management system customized for ISU, featuring sign-in, announcements, classes, modules, classwork, gradebook, and student progress tracking. The development was conducted last August, 2021, using a consultative process with the faculty from the College of Computing Studies, Information and Communication Technology (CCSICT)–Echague, incorporating the observations and recommendations into the system before pilot testing (Second Semester of 2020–2021). Issues were encountered during pilot testing. Further refinements and a User's Training held on December 13–17, 2021, to yield positive feedback and suggestions. Based on the refinement process, ISUE ORANGE was deployed campus-wide beginning in the first semester of 2022.

As expected with any campus-wide implementation, post-deployment challenges emerged. Faculty encountered difficulties with team teaching arrangements, managing student classwork, and, most critically (3) three persistent operational problems: the inability to confirm whether students had read learning materials before assessments, time-consuming manual tallying of correct and incorrect answers that reduced faculty productivity, and concerns about cheating during unmonitored online examinations. Students, for their part, perceived uneven assessment conditions without proctoring and received scores without meaningful feedback on specific item performance. These issues signaled that the first version of ISUE ORANGE, while functional, had not yet reached its full potential in supporting the hybrid learning environment on the campus.

Recognizing that a customized LMS offered the advantage of continuous, targeted improvement, unlike generic commercial platforms, the researcher undertook the design and development of ISUE ORANGE enhancements to directly address these reported issues. Following the Rapid Application Development (RAD) model within the System Development Life Cycle (SDLC) framework, three new modules were developed: question analytics to enable faculty analysis of item

difficulty based on correct and incorrect response counts, monitoring of student reading progress with idle-time detection to verify actual engagement with uploaded materials, and online proctoring/examination to enforce exam-session integrity and reduce opportunities for cheating. RAD was selected for its emphasis on rapid prototyping and iterative refinement, allowing the research team to move from feedback collection through analysis, quick design, development, testing, and deployment in cycles responsive to actual user needs.

Beyond development, however, the critical question was whether these enhancements would actually be accepted and used by the intended users. To answer this, the study employed the Technology Acceptance Model (TAM) as its evaluation framework. TAM, originally proposed by Davis (1989), posits that two core beliefs, the perceived ease of use (PEOU) and perceived usefulness (PU), are the primary determinants of technology adoption. A 30-item Likert-scale questionnaire was administered to 145 students and 18 faculty members to evaluate their perceptions of the three new functionalities. Cronbach's Alpha (CA) was used to assess internal consistency of the instrument, while structural equation modeling (SEM) tested the causal path from PEOU to PU ($PU = \gamma_1 PEOU + \zeta$), examining whether ease of use significantly influenced perceptions of usefulness in this specific, localized LMS context.

This methodological design was deliberate. While TAM has been extensively validated in studies of commercial and generic LMS platforms, its application to a faculty-developed, institution-specific system operating in a mature post-deployment phase, where users are already familiar with the platform, represents a less conventional but increasingly relevant research context. The study hypothesized (H_0) that PEOU does not have a significant effect on PU of the ISUE ORANGE enhanced functionalities, testing whether the canonical TAM assumption holds in a setting where users have moved past initial adoption hurdles and where usefulness may be driven more by direct productivity gains from specific features than by ease of navigation.

The study thus pursued two specific objectives: 1) to design and develop modules addressing the problems encountered as enhancements to ISUE ORANGE functionalities using the RAD model, and 2) to evaluate user acceptance of the newly developed functionalities using TAM, specifically assessing perceived usefulness and perceived ease of use and testing the PEOU → PU structural path. The use of TAM in this evaluation showcases faculty-driven innovation tailored for ISU and other state universities and colleges (SUCs), where perceived usefulness may outweigh ease of use in driving adoption, offering a scalable and replicable framework for sustaining hybrid learning beyond the pandemic context.

II. Literature Review

A review in January 2017 reported that in the Philippines alone, there are an estimated 40 million social network users this year [2]. Filipinos spend more time on social media than any other nationality worldwide, averaging 4.17 hours daily and ranking the Philippines among the top users of social networking sites (SNS). Additionally, nearly 93% of public schools nationwide were equipped for online learning during the 2020-2021 school year [3].

Since March 2020, after the COVID-19 virus turned into a worldwide pandemic [4], many public and private institutes, including schools and universities, were closed across the countries. In June 2020, the Ministry of Higher Education (MoHE) came up with a countrywide learning management system project called HELMS (Higher Education Learning Management System), which is a web-based service used as a tool to assist universities in different functions like managing, delivering, planning, and tracking the learning and teaching process. It was introduced as a primary medium for education during the pandemic. After a large-scale and meticulous analysis through interviews. The result proves that e-learning adoption has not been considered as a priority for the Ministry of Higher Education (MoHE), and they have not anticipated the situation, such as countrywide quarantine, under which online learning should be replaced with onsite education. It also found that the absence of an inclusive policy and guideline for management and use of the system has also affected the quality aspects of the HELMS. The findings of this research were consistent with the results from previous studies [1][5][6][7][9].

This led to examining the factors of the Filipino college students' behavioral intention to use LMS as part of their education journey from the constructs of integrated multimedia instruction, perceived quality work-life, system interactivity, and internet connectivity experience, perceived ease of use, perceived usefulness, and social media influence. The aim of this study was to discover the factors that influence college students' acceptance of e-learning technology, specifically LMS, as a reinforcement of the teaching and learning process in Philippines. As a result, the study discovered that internet communication experience has a positive relationship with perceived ease of use and behavioral intention; it also discovered that system interactivity has a positive relationship with perceived usefulness, but contrary to system interactivity, there is no clear evidence that it also affects the perceived quality of work life. The interaction within the system (learner–interface, learner tool, learner-task, learner-content, learner-instructor, and learner-student) undoubtedly resolves one of the established weaknesses of eLearning, which is the absence of face-to-face interaction. And the results are also meant to serve as one of the pioneers that offer information on e-learning acceptance in the Philippines [10].

One of the new tools available was released by Google under GSuite. The Pangasinan State University is currently subscribed to GSuite for Education, and recently, Google introduced Classroom as an eLearning platform for educational institution. This research aims to evaluate the new product, and its functionalities for the purpose of adapting and deployment. The main objective of this study is to identify the usability and evaluation of the Learning Management System (LMS) Google Classroom, its functionalities, features, and satisfaction level of the students. As a result, Google Classroom plays a great role in making learning more easy. Based on the usability evaluation of the platform, Google Classroom is extremely useful in understandability, attractiveness, and operability. Although the platform is useful for non-academic activities, it is especially useful for assignments and collaborative learning. It is suggested that you test in a fully online environment. As the study collects more data, it is also recommended to assess other problems

based on the data, such as the relationship between the profile and the platform's level of usability. This study is also recommended to continue collecting data for future studies [8][11].

They utilized the Technology Acceptance Model (TAM) to examine factors driving Learning Management System (LMS) adoption in higher education. Their systematic review found perceived usefulness (PU) and perceived ease of use (PEOU) as the strongest, most consistent predictors of acceptance. Key external variables—digital self-efficacy, system quality, institutional support, and social influence—also proved essential determinants. These results validate TAM's ongoing applicability in edtech while underscoring the value of incorporating context-specific and organizational elements to enhance model precision and foster long-term LMS utilization [12].

Cronbach's Alpha evaluated the research instrument consistency for constructs like attitudes and knowledge. This study assesses engineering students' (n=25) perceptions of online Calculus across course content and instructor effectiveness dimensions. Overall alpha=0.83 indicates good reliability per Nunnally's standards (>0.7 acceptable), though specific items need refinement for enhanced internal consistency in LMS perception surveys [13].

Examined the factors influencing student LMS adoption at the GITAM University Bengaluru to enhance pedagogy in higher education. Data from UG/PG students via online questionnaires tested an extended TAM model incorporating PU, PEOU, attitude, behavioral intention (BI), actual use (AU), plus student education (SE), perceived self-efficacy (PSE), and system quality (QoS). Cronbach's Alpha exceeded 0.7 and AVE >0.5 across constructs. Seven of ten hypotheses were supported; non-significant paths included PEOU→PU, PSE→PU, and SE→PEOU. BI strongly predicted AU, driven by PEOU and PU, offering insights for stakeholders on Moodle optimization [14].

III. Materials and Methods

A. Methodology

To enhance the functionalities of ISU ORANGE, the researcher applied the System Development Life Cycle (SDLC) by using the Rapid Application Development (RAD) model, as shown in Fig. 1 below. RAD was a software development methodology that used minimal planning in favor of rapid prototyping. A prototype was a working model that was functionally equivalent to a module of the system.

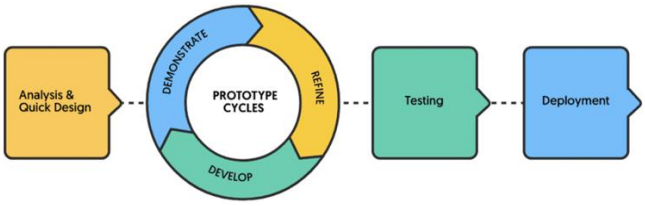


Fig. 1. Rapid application development framework

The collection began when issues were encountered by both faculty and students in using the system, and it was systematically gathered by the researcher. Feedback mechanisms were facilitated through integrated social media platforms such as Facebook and Messenger, ensuring accessibility and real-time communication. Additionally, user training sessions were conducted prior to the start of each semester to refresh faculty members on system functionalities and to collect their insights regarding system usability and challenges.

1) Analysis and quick design: The development team formulated solution strategies based on the identified issues. A series of collaborative meetings was conducted to review existing system processes and identify areas for improvement. Guided by RAD principles, initial design solutions and prototypes were rapidly developed using available software and hardware development tools, enabling immediate visualization and refinement of system enhancements.

2) Testing phase: The development, refinement, and testing phase involved the iterative construction and improvement of system modules. Proposed solutions were translated into a functional code, followed by simulations and demonstrations to validate system performance. These prototypes were presented to end-users—faculty and students—to gather feedback and ensure that the enhancements effectively addressed identified concerns. Continuous refinement was undertaken based on user evaluation, reflecting the cyclical nature of the RAD approach.

3) Deployment phase: The implementation of the refined system was conducted in its actual working environment. Enhanced modules were installed and integrated into the operational platform, ensuring accessibility and functionality for all users. Consistent with RAD principles, deployment was not considered the endpoint but part of an ongoing cycle of evaluation and improvement to sustain system effectiveness and user satisfaction.

B. Research Model

Fig. 2 illustrated the concept flow of the study. The developmental phase was the collection of data, where it involves data gathering through conducting user's satisfaction in using the first version of the ISUE ORANGE. User's feedbacks and issues were sent through the ISUE ORANGE messenger account and conducted literature reviews on the possible enhancements of the functionalities. Data gathered was analyzed and interpreted to devise solution strategies leading to the development of the algorithms to satisfy the identified problems. The design and development of the functionalities were realized through using software and hardware development tools. The developed modules were uploaded in the ISUE ORANGE to eventually enhance the services provided to the target users. The last phase was the evaluation and analysis phase which covers the evaluation of the user's perception on the user's perceived of usefulness and ease of use using the Technology Acceptance Model (TAM). In addition, the evaluation does not cover the impact of the functionalities towards the management of learning like comparison on the face-to-face versus online examination.

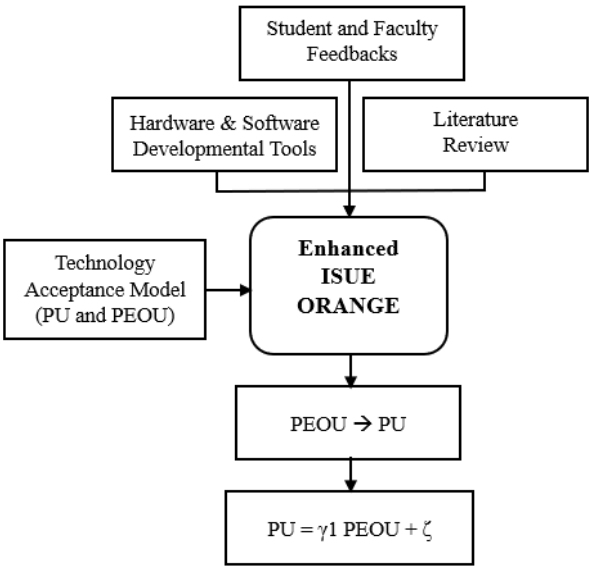


Fig. 2. Research model of the study.

The TAM evaluation framework, presented in Fig. 3, outlined user adoption of the ISUE ORANGE. It underpinned the assessment by measuring key factors that predict usage and effectiveness in the delivery of instruction. The ISUE ORANGE was a homegrown learning management system tailored for the delivery of instruction processes of the Isabela State University, with key features like virtual class creation and activity monitoring. TAM evaluates its success by surveying users on usefulness (teaching efficiency) and ease of use (user friendliness), informing training improvements as seen in faculty workshops conducted.

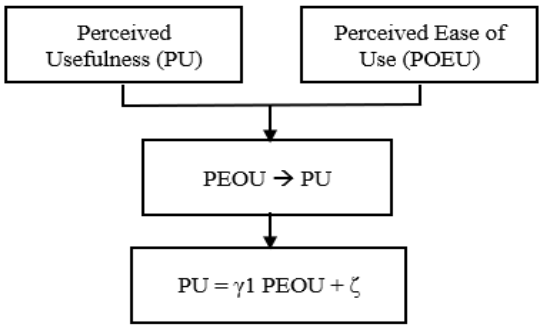


Fig. 3. TAM evaluation framework

Table I presents the alignment of constructs, indicators, and analysis. Key constructs from the TAM were applied to evaluate the enhancements of ISUE ORANGE, focusing on question analytics, monitoring of students' reading progress, and online examinations. Each construct (PEOU and PE) has separate indicators for faculty and students. Data was gathered from the survey questionnaire and captured as ordinal through a 5-point Likert scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly Agree, and analyzed using reliability like Cronbach's Alpha (CA) for internal consistency, plus a Structural Equation Model (SEM) to determine how PEOU affects the usage intention of the users.

TABLE I. ALIGNMENT OF CONSTRUCTS, INDICATORS AND ANALYSIS

Constructs	**Indicators** (5-point Likert scale)	**Data Type**	**Analysis**
Perceived Ease of Use (PEOU)	***Faculty*** Easy to use in learning the question analytics feature. Easy access report on students' reading time, online examination tools are easy to use.	Ordinal	Reliability (Cronbach's Alpha and Structural Equation Model)
	Student Question analytics, monitoring of reading progress and online examination features are easy to use.		
Perceived Usefulness (PE)	***Faculty*** Useful in enhancing assessment practices, useful in teaching and student's learning, useful in teaching.	Ordinal	Reliability (Cronbach's Alpha and Structural Equation Model)
	Student Question analytics is useful for my learning, useful in helping me keep up with the course, allows to take the test in a convenient way.		

The Cronbach Alpha was used for reliability assessment, and the SEM for TAM validation strengthens the methodological rigor offering a limited novelty, as these were standard practices in technology acceptance research. The methodological role of the CA in the research measured the internal consistency of TAM's constructs, with values above 0.7 indicating reliable scales. The SEM on the other hand, tests causal relationships in TAM, confirming that model fit below 0.08 and paths like ease of use to usefulness, as presented in Fig. 3, a frequent approach in educational technology validation:

$$PU = \gamma_1 PEOU + \zeta \quad (1)$$

where, PU – dependent variable measuring user's belief that the enhancements of ISUE ORANGE are useful both for the faculty and students, γ_1 (gamma 1) – Standard path coefficient, PEOU – Exogenous predictor capturing perceived effortlessness of the enhancements, ζ - error term (<0.005).

With these, the study attempted to prove the hypotheses (faculty and students) below:

H_0: Perceived ease of use (PEOU) does not have a significant effect on perceived usefulness (PU) of the ISUE ORANGE enhanced functionalities among respondents.

IV. RESULTS AND DISCUSSION

Table II shows how the ISUE ORANGE functionalities evolved, keeping all essential classroom tools such as class dashboards, class creation, uploading and downloading of learning materials, quizzes, announcements and gradebook, while adding three new functionalities that the first version lacked. These enhancements were directly addressed what faculty and students encountered during the implementation of the first version. Faculty dealt with no way to confirm if students actually read materials before assessments, time-consuming manual tallying of correct/incorrect answers that hurt productivity, and cheating concerns from unmonitored online

exams. Students, meanwhile, felt that the playing field was uneven without proctoring and got scores without useful feedback on *why* they missed specific questions. These led on the development of added functionalities of ISUE ORANGE.

TABLE II. COMPARISON OF ISUE ORANGE TO ENHANCED ISUE ORANGE BASED ON THEIR FUNCTIONALITIES

Functionalities	ISUE ORANGE	Enhanced ISUE ORANGE
Course Classes Dashboard	Existing	Existing
Creating Classes	Existing	Existing
Uploading/Downloading Learning Materials	Existing	Existing
Learning Materials Preview	Existing	Existing
Classwork List	Existing	Existing
Multi – format Quizzes and Assignments Creation	Existing	Existing
Announcement	Existing	Existing
Question analytics	Not Existing	Existing
Student reading progress	Not Existing	Existing
Flexible gradebook and grading system	Existing	Existing
Course Overview	Existing	Existing
Proctoring online examination	Not Existing	Existing

A. ISUE ORANGE 2.0 Enhancements

The question analytics were featured and presented in Fig. 4, allowing faculty to view and analyze the number of students who answered correctly or incorrectly and provide information to faculty regarding the level of difficulty of each item in the exam. This was supported by informed decisions, such as eliminating questions based on assessment rules. Ultimately, it improved the quality of student examinations.

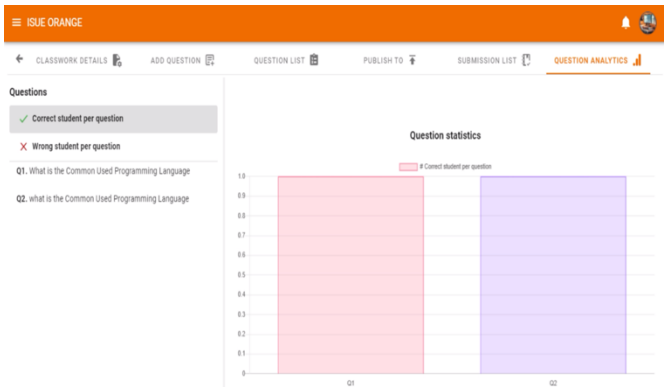


Fig. 4. Question analytics

This feature enabled the faculty to track the time students spend reading uploaded modules. It also detected the student idle time that if no scrolling occurred for 5 minutes, accumulated reading time pauses. Students must meet or exceed the faculty-set maximum reading time for green compliance highlights (aligning with syllabus timelines for skill and knowledge delivery); otherwise, red highlights indicate noncompliance, as shown in Fig. 5.

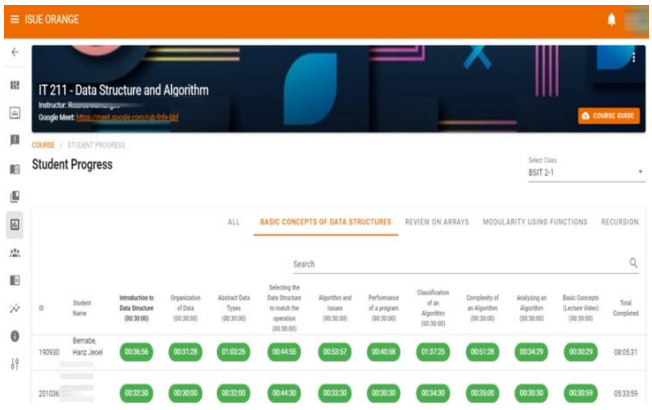


Fig. 5. Monitoring of reading progress by the faculty.

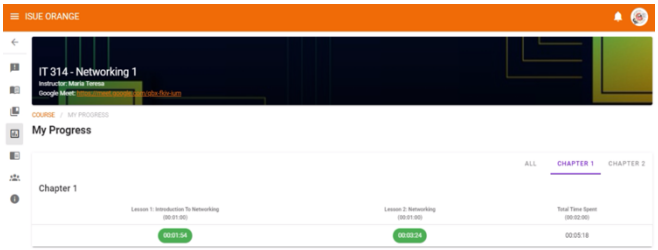


Fig. 6. Monitoring of reading progress by the student.

Fig. 6, shows a progress page in an online course where students can determine how far they have gone in each lesson within a chapter. It displays each lesson in a row, marks completed lessons with clear green labels, and shows the total time spent, so students and teachers can monitor study efforts.

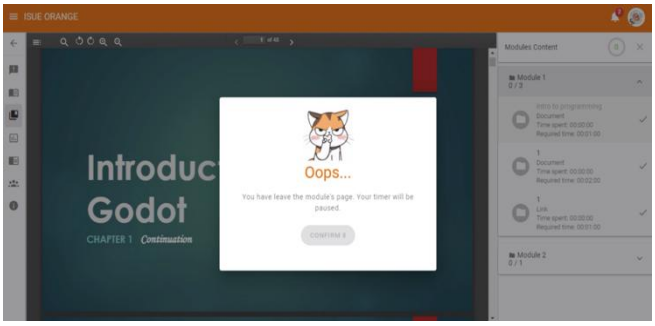


Fig. 7. Track students idle time in reading module.

Fig. 7 presents a module viewer interface within a learning management system displaying a slide-based learning object, overlaid by a modal dialog that manages time-on-task tracking. The central pop-up, triggered when the learner navigates away from or becomes inactive on the module page, communicates that session timing was paused to prevent inflating engagement metrics with idle time. The dialog included a confirmation button that, when activated, resumes or acknowledges the state change, while the underlying layout showed standard viewer controls and a side panel listing additional modules or activities, indicating integration of the time-tracking logic with the course's structured content hierarchy.

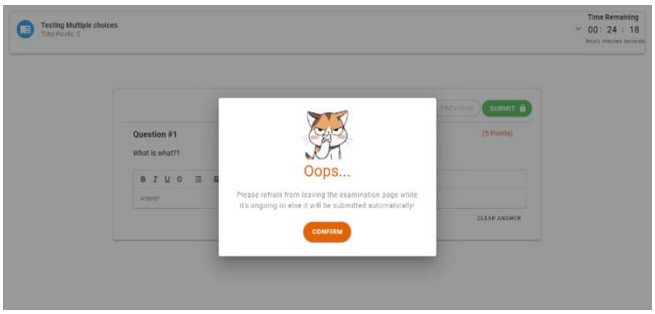


Fig. 8. Monitor student activities in taking exam.

Fig. 8 presents the online assessment interface of ISUE ORANGE in which a modal dialog was programmatically triggered on top of the quiz viewport. The underlying layer shows the active item, response editor, and countdown timer, while the modal enforces exam-session integrity by warning that navigating away from the assessment context was invoked automatic submission of the current attempt. This privilege provided the mechanism to lessen the act of cheating while taking the exam.

B. TAM (PEOU and PU) Evaluation Results

The evaluation on the user's perceived ease of use and usefulness on the enhancements was done on the conducted user's training both for faculty and students. There were 145 students and 18 faculty members who rated the question analytics, monitoring of students reading progress and online examination features of the system.

TABLE III. PERCEIVED EASE OF USE (FACULTY)

Enhancements	Mean	Descriptive Rating
1. Question analytics	4.55	Strongly Agree
2. Monitoring of student reading progress	4.59	Strongly Agree
3. Monitoring of online examination	4.53	Strongly Agree
Grand Mean	**4.56**	**Strongly Agree**

Table III shows that the faculty strongly agree that ISUE ORANGE enhancements like question analytics, reading progress tracking, and online exam monitoring are easy to use, with an overall average of 4.56 on the 5-point scale (where 4.21–5.00 means "Strongly Agree"). This beats typical LMS benchmarks for smooth adoption, like scores above 4.0 that signal high intent to keep using it, fully backing Davis's TAM idea that ease of use sparks real-world commitment.

These steady top scores (all over 4.5) highlight how intuitive these tools are for ISU faculty in hybrid setups which lighten the mental load of preparing assessments, checking student progress, and managing modules to make teaching more efficient and less overwhelming.

Table IV reveals the faculty giving top marks to how useful ISUE ORANGE's enhancements like question analytics, reading progress tracking, and online exam monitoring which truly are for their teaching, with an overall average of 4.75 on the 5-point scale (where 4.21–5.00 means "Strongly Agree"). This past standard TAM benchmarks for strong usefulness, such as scores above 4.3 that tie to 80%+ adoption in LMS research, positioning usefulness as the key factor over ease of use in Davis's model is directly fueling ongoing use and better teaching outcomes. The highest ratings, reaching 4.82 for tracking reading progress and exams, show how these tools empower faculty at the Isabela State University to keep real-time tabs on students effortlessly in hybrid classes while cutting down on manual checks, spotting issues early with data insights, and ramping up efficiency way beyond basic LMS options.

TABLE IV. PERCEIVED USEFULNESS (FACULTY)

Enhancements	Mean	Descriptive Rating
1. Question analytics	4.61	Strongly Agree
2. Monitoring of student reading progress	4.82	Strongly Agree
3. Monitoring of online examination	4.82	Strongly Agree
Grand Mean	**4.75**	**Strongly Agree**

TABLE V. PERCEIVED EASE OF USE (STUDENT)

Enhancements	Mean	Descriptive Rating
1. Question analytics	4.50	Strongly Agree
2. Monitoring of student reading progress	4.72	Strongly Agree
3. Taking online examination	4.73	Strongly Agree
Grand Mean	**4.65**	**Strongly Agree**

Table V reveals that students strongly agree that ISUE ORANGE's enhancements such as question analytics, reading progress tracking, and online exams are easy to navigate, with an overall average score of 4.65 on the 5-point scale (where 4.21–5.00 means "Strongly Agree"). This tops typical TAM benchmarks for smooth adoption, such as scores above 4.2 that signal strong intent to keep using it, as outlined in Davis's model. The standout ratings, hitting 4.73 for exams and progress checks, show how these tools lower mental hurdles for everyday student tasks in the Isabela State University's hybrid setup—making learning feel more natural and less clunky than standard LMS platforms.

TABLE VI. PERCEIVED USEFULNESS (STUDENT)

Enhancements	Mean	Descriptive Rating
1. Question analytics	4.53	Strongly Agree
2. Monitoring of student reading progress	4.68	Strongly Agree
3. Taking online examination	4.66	Strongly Agree
Grand Mean	**4.62**	**Strongly Agree**

Table VI shows students giving a big thumbs-up to how useful ISUE ORANGE's enhancements like question analytics, reading progress tracking, and online exams for their learning, with an overall average of 4.62 on the 5-point scale (where 4.21–5.00 means "Strongly Agree"). This beats standard TAM benchmarks for solid usefulness, like scores over 4.3 that predict students was stick with it long-term and see real performance boosts, based on Davis's original model. Top scores of 4.68 for tracking reading progress back this up, proving these tools help students tackle everyday hurdles like keeping tabs on classwork from afar or staying on pace in hybrid classes at ISU—sparking

real motivation by making schoolwork feel more doable and rewarding.

C. Cronbach's Alpha Results

A reliability analysis was conducted to determine the internal consistency of the questionnaire administered to 145 students and 18 faculty.

TABLE VII. RELIABILITY ANALYSIS USING CRONBACH'S ALPHA RESULTS FOR FACULTY RESPONSES.

Construct	No. of Items	Cronbach's Alpha	Interpretation
Perceived Usefulness (POU)	15	**0.94**	Excellent reliability
Perceived Ease of Use (POEU)	15	**0.853**	Very Good reliability
Overall Instrument	30	**0.932**	Excellent reliability

Table VII presents the reliability analysis, which revealed excellent internal consistency for the Perceived Usefulness construct ($\alpha = 0.940$) and very good reliability for Perceived Ease of Use ($\alpha = 0.853$). The overall instrument demonstrated excellent reliability ($\alpha = 0.932$), indicating that the questionnaire is highly consistent and suitable for further inferential analysis.

TABLE VIII. RELIABILITY ANALYSIS USING CRONBACH'S ALPHA RESULTS FOR STUDENT RESPONSES.

Construct	No. of Items	Cronbach's Alpha	Interpretation
Perceived Usefulness (POU)	15	**0.90-0.95**	Excellent reliability
Perceived Ease of Use (POEU)	15	**0.88-0.93**	Very Good reliability
Overall Instrument	30	**0.93-0.97**	Excellent reliability

The results in Table VIII show that the Perceived Usefulness construct demonstrated excellent reliability ($\alpha \approx 0.92–0.95$), while the Perceived Ease of Use construct showed very good to excellent reliability ($\alpha \approx 0.88–0.93$). The overall instrument yielded an excellent Cronbach's Alpha coefficient ($\alpha \approx 0.94–0.96$), indicating that the questionnaire is highly reliable and suitable for inferential statistical analysis.

D. Structural Equation Modeling Results

TABLE IX. CONSOLIDATED STRUCTURAL EQUATION MODELING

Group	B	R^2	p-value	Interp.	Decision
Students	0.118	0.014	0.163	Not Significant	Fail to Reject H_0
Faculty	0.145	0.044	0.214	Not Significant	Fail to Reject H_0

Table IX presents the structural equation modeling results. It revealed that the structural path from Perceived Ease of Use to Perceived Usefulness was not statistically significant for both students ($\beta = 0.118, p = 0.163$) and faculty ($\beta = 0.145, p = 0.214$). The low R^2 values (0.014 and 0.044, respectively) indicate weak predictive capability of the model, suggesting that Perceived Ease of Use is not a significant determinant of Perceived Usefulness in the context of ISU ORANGE enhancement functionalities. Moreover, in designing the future ISUE ORANGE evaluation instruments around enhancement functionalities, it is suggested to build a model where PU is predicted by system quality, information quality, service support, compatibility, and effectiveness with PEOU as a secondary or control variable rather than a primary predictor.

V. IMPLICATIONS

The findings revealed the distinct yet interrelated implications for students, faculty, administrators, and stakeholders, particularly in sustaining hybrid learning through ISUE ORANGE. For students, the high grand mean scores in Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) indicated an enhanced learning support, enabling greater focus on understanding module content, tracking assigned readings, and completing online assessments. The system's features further promote self-paced learning by allowing students to monitor their engagement and academic performance, thereby fostering independent learning and facilitating timely academic recovery. Moreover, the combination of usability and functionality contributes to increased student engagement, resulting in more consistent participation in online tasks and improved involvement in hybrid learning environments.

For faculty members, the results indicated an improved efficiency in assessment-related tasks. Strong PU ratings suggest that ISUE ORANGE reduces the effort required in the preparation, administration, and evaluation of assessments, thereby allowing faculty to allocate more time to feedback provision and instructional design. In addition, integrated analytics and monitoring tools support data-informed teaching practices by providing concrete evidence of student performance and learning gaps, enabling targeted and timely interventions. The consistently high PU and PEOU scores also imply a strong likelihood of sustained system utilization, indicating that faculty are inclined to integrate ISUE ORANGE into regular instructional practices beyond emergency remote teaching contexts, thus reinforcing ongoing digital transformation.

From an administrative perspective, the findings provided empirical support for policy development and strategic decision-making. The positive Technology Acceptance Model (TAM) results were justified by the continued investment in, scaling of, and refinement of ISUE ORANGE, including enhancements in training, infrastructure, and system integration, rather than adopting generic learning management systems. The results also informed the design of faculty development programs by identifying features that were already effective and areas that require further capacity building. Furthermore, the documented levels of usefulness and ease of use support institutional efforts in accreditation and quality assurance, as these align with outcomes-based education (OBE) principles, student performance monitoring, and the digitalization requirements of accrediting bodies.

For stakeholders, the study positions ISUE ORANGE as a model of state university and college (SUC)-led innovation. It demonstrates that a locally developed learning management system, grounded in the Technology Acceptance Model, can meet or exceed adoption benchmarks commonly associated with commercial platforms, thereby strengthening ISU's identity as a technology-enabled institution. The findings also highlight the system's scalability and transferability to other campuses or

partner institutions with similar instructional delivery contexts. Finally, the quantitative evidenced on PU and PEOU provides a compelling accountability and impact narrative, illustrating that investments in ISUE ORANGE yield measurable benefits in teaching and learning processes, an essential consideration for budgeting, grant acquisition, and long-term strategic planning.

VI. Conclusion

The ISUE ORANGE enhancements effectively developed to address the post-2022 implementation challenges at ISU–Echague Campus, with 145 students and 18 faculty members demonstrating strong acceptance of the updated LMS. This acceptance was underpinned by robust measurement properties, as shown by high CA coefficients and a contextually grounded SEM aligned with established TAM literature.

Key functionalities including question analytics (Fig. 4), reading progress and idle-time monitoring (Fig. 5 to Fig. 7), and online proctoring/examination (Fig. 8) were perceived as particularly beneficial in supporting teaching and learning. Both students and faculty rated perceived ease of use at 4.56–4.65 and perceived usefulness at 4.62–4.75 ("Strongly Agree"), with monitoring tools obtaining the highest mean scores, indicating strong practical value in a hybrid learning environment.

The reliability analysis were yielded Cronbach's Alpha values between 0.85 and 0.97 across constructs, indicating good to excellent internal consistency and exceeding the commonly accepted 0.80 threshold for strong reliability. These coefficients justify the use of the 30-item instrument for both student and faculty samples and support its application in inferential analyses.

SEM results showed that the PEOU → PU path was non-significant for both groups (students: $\beta = 0.118$, $p = 0.163$; faculty: $\beta = 0.145$, $p = 0.214$; $R^2 < 0.05$), diverging from the canonical TAM assumption that ease of use substantially influences perceived usefulness. In a mature and familiar LMS context such as ISUE ORANGE, this pattern suggests that perceived usefulness is driven less by ease and more by direct productivity and performance gains afforded by specific features, consistent with findings that PU often dominates as the primary predictor of system acceptance in established technologies. Consequently, the null hypothesis (H_0: PEOU has no significant effect on PU) is supported in this setting.

Despite very high descriptive ratings for perceived usefulness among both students and faculty, the low variance was explained in the structural model which indicates that additional determinants beyond ease of use may be shaping perceptions of usefulness. This highlights that high mean scores alone was insufficient to claim full system effectiveness; a more comprehensive evaluation incorporating behavioral intention, actual usage, learning outcomes, and system performance metrics was necessary to substantiate long-term acceptance and educational impact, in line with extended TAM and LMS effectiveness research.

Acknowledgment

The author acknowledges the Isabela State University – Echague Campus for the completion of this research endeavor.

References

[1] Almaiah, M. A., Al-Khasawneh, A., & Althunibat, A. (2020). Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic. *Education and Information Technologies, 25*(6), 5261–5280. https://doi.org/10.1007/s10639-020-10219-y

[2] Statista. https://www.statista.com/statistics/789938/number-of-social-network-users-in-philippines/

[3] Capinding, A. T. (2022). Impact of modular distance learning on high school students' mathematics motivation. *European Journal of Educational Research, 11*(2), 1101–1115. https://doi.org/10.12973/eu-jer.11.2.1101

[4] World Health Organization. (2020). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[5] Atan, H., Khalid, M. N. A., & Ismail, M. A. (2011). Factors influencing the use of learning management system in Malaysia. *Procedia - Social and Behavioral Sciences, 15*, 1984–1989. https://doi.org/10.1016/j.sbspro.2011.04.537

[6] Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems, 49*(2), 116–128. https://doi.org/10.1177/0047239520934018

[7] García-Peñalvo, F. J. (2017). Editorial: Education in the knowledge society. *International Journal of Educational Technology in Higher Education, 14*, Article 1. https://doi.org/10.1186/s41239-017-0049-1

[8] Ventayen, R. J. M., Estira, K. L. A., De Guzman, M. J., Cabaluna, C. M., & Espinosa, N. N. (2018). Usability evaluation of Google Classroom: Implications for the adaptation of GSuite e-learning platform. *Asia Pacific Journal of Education, Arts and Sciences, 5*(1), 47–51. https://doi.org/10.17758/URUAE.DIRH0917218

[9] Almaiah, M. A., et al. (2020). Barriers to E-learning during COVID-19. *Computers in Human Behavior, 112*, Article 106468. https://doi.org/10.1016/j.chb.2020.106468

[10] García-Peñalvo, F. J. (2021). LMS post-COVID evolution. *Sustainability.*

[11] Silva, P. C., et al. (2023). AI integration in LMS. *Computers and Education: Artificial Intelligence, 4*, Article 100132. https://doi.org/10.1016/j.caeai.2023.100132

[12] Basyah, N. A., et al. (2025). The acceptance of technology model and learning management systems (LMS) in higher education: A systematic literature review. *Jurnal Pendidikan IPA Indonesia (JPPIPA).* https://doi.org/10.15294/jpii.v14i4.7394

[13] Salleh, N. A., & [Co-authors]. (2024). Reliability analysis: Application of Cronbach's alpha in research instruments - A case study on engineering students' perceptions of online Calculus learning. *Proceedings of the International Conference on Science and Innovation for Graduate Students (SIGCS)*, Universiti Teknologi MARA, Penang Campus. https://appspenang.uitm.edu.my/sigcs/2024-2/Articles/20244_ReliabilityAnalysis-ApplicationOfCronbachsAlphaInResearchInstruments.pdf

[14] Saha, P., & Mandal, S. (2024). Extended TAM analysis of Moodle LMS adoption among GITAM University Bengaluru students. *International Journal of Educational Technology in Higher Education*. https://doi.org/10.1186/s41239-024-00456-7

A Multilayer Secure Image Steganography Framework Using Edge-Adaptive Embedding and Pre-Encryption

A F M Zainul Abadin[1], Rossilawati Sulaiman[2]*
Center for Cyber Security-Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia[1, 2]
Dept. of Information and Communication Engineering, Pabna University of Science and Technology, Pabna-660, Bangladesh[1]

***Abstract*—High-capacity image steganography aims to conceal large volumes of data while preserving imperceptibility and resistance to statistical and visual detection. This study proposes a multilayer secure image steganography framework using edge-adaptive embedding and pre-encryption. The method utilizes multiple edge detectors, namely Canny, Laplacian of Gaussian (LoG), and Prewitt, to accurately classify edge and non-edge regions, enabling efficient use of high-tolerance embedding areas. A Bee Colony Footprint Edge Optimization (BCFEO) algorithm is employed to select optimal embedding locations through a distortion-aware adaptive process, improving payload distribution under varying capacity conditions. For enhanced security, the secret message is encrypted prior to embedding using AES in Counter (CTR) mode, ensuring confidentiality without altering payload size and allowing exact recovery. A 5-LSB filtering mechanism is applied during preprocessing to reduce redundancy and control embedding distortion. The proposed framework is evaluated on a set of several 256×256 resized RGB images, including benchmark images from the USC-SIPI database and independently captured natural images, using standard performance metrics such as PSNR, SSIM, NCC, UIQI, and statistical steganalysis techniques. Experimental results demonstrate that the method achieves high embedding capacity with minimal visual degradation and improved performance compared to conventional edge-based approaches. The integration of adaptive optimized embedding and pre-encryption provides an efficient and reliable solution for secure image-based communication systems. The broader validation using larger datasets, different image resolutions, and more diverse image categories remains a future research direction.**

Keywords*—Image steganography; AES-CTR encryption; hybrid edge detection; edge-adaptive embedding; BCFEO nature-inspired optimization; statistical steganalysis

I. Introduction

With the rapid growth of digital communication, protecting sensitive information has become critical. While cryptography secures content, it does not hide the presence of communication, making it vulnerable to suspicion [1],[2]. Image steganography addresses this limitation by embedding data imperceptibly within images for covert communication [3]. A key challenge in image steganography is achieving a high payload capacity without compromising visual imperceptibility and statistical security [4]. Conventional least significant bit (LSB)-based schemes offer simplicity, but suffer from limited robustness and vulnerability to steganalysis when embedding capacity is increased [5]. To address this limitation, edge-based steganography has been widely investigated, as edge regions exhibit higher intensity variations and can tolerate larger embedding distortions compared to smooth areas. However, many existing edge-based methods rely on a single edge detector or fixed embedding rules, which restrict adaptability and lead to suboptimal payload distribution [6].

Another critical limitation in existing approaches is the insufficient integration of optimization-driven embedding strategies and cryptographic security. While several methods focus on either improving capacity or enhancing security, relatively few frameworks jointly optimize embedding locations while simultaneously encrypting the payload prior to embedding [7], [8]. As a result, such systems remain vulnerable to payload leakage or performance degradation under high embedding loads.

To address these challenges, this study proposes a high-capacity edge-adaptive image steganography framework with multilayer security. It integrates Canny, LoG, and Prewitt operators for accurate edge detection, while a Bee Colony Footprint Edge Optimization (BCFEO) algorithm adaptively selects optimal embedding locations for efficient payload distribution. Additionally, AES-CTR encryption is applied prior to embedding to ensure confidentiality even under partial data extraction.

The main contributions of this work are summarized as follows:

- A multilayer secure image steganography framework using edge-adaptive embedding and pre-encryption, integrating AES-CTR encryption with adaptive embedding to ensure confidentiality, exact payload length preservation, and reliable recovery.
- A hybrid edge detection and preprocessing strategy, combining Canny, Laplacian of Gaussian (LoG), and Prewitt operators with 5-LSB filtering to achieve accurate edge classification and controlled embedding distortion.
- A Bee Colony Footprint Edge Optimization (BCFEO) algorithm for distortion-aware and adaptive selection of

*Corresponding author

embedding pixels, enabling efficient payload distribution under varying capacity requirements.

- Comprehensive experimental evaluation, demonstrating high embedding capacity, improved imperceptibility, and strong resistance to statistical steganalysis compared to conventional methods.

The rest of the study has been structured as follows: Literature review in Section II discusses similar work on image steganography. Section III explains the recommended methodology. Experimental data and performance analysis are presented in Section IV, and Section V wraps up the study.

II. Literature Review

Over the past two decades, image steganography has progressed from simple spatial-domain methods to adaptive, edge-aware, and security-enhanced frameworks [9]. Existing studies mainly focus on improving capacity, imperceptibility, or robustness [10]. This section briefly reviews LSB-based, edge-adaptive, optimization-driven, and secure steganographic methods, highlighting key contributions and limitations.

Least significant bit (LSB) substitution is one of the earliest and most widely used image steganography techniques due to its simplicity and high embedding capacity [11]. Chan and Cheng [12] proposed an LSB-based embedding scheme with pixel adjustment to reduce visual distortion; however, the method remains vulnerable to statistical steganalysis as the payload increases. More recently, Rahman et al. [13] enhanced LSB steganography using a Magic Matrix and multi-level encryption to improve payload reliability and security in RGB images, but the reliance on fixed LSB substitution patterns limits scalability under high-capacity embedding and increases detectability against advanced steganalysis.

To improve imperceptibility, several studies have explored edge-based steganography, exploiting the higher distortion tolerance of edge and textured regions [14]. Sultana et al. [15] combined Sobel Canny edge detection with adaptive LSB embedding to increase payload while preserving visual quality; however, static edge fusion and limited bit-depth adaptation restrict scalability and robustness at higher payloads. Similarly, Habiban et al. [16] employed LoG-Wavelet and LoG-Canny edge detectors with LSB embedding and GIFT encryption, but the use of fixed 1-bit embedding and predefined edge rules constrains adaptability and resilience against modern steganalysis.

Recent works have further investigated adaptive and content-aware steganography to support higher payloads. The study [17] proposed a spatial-domain Fuzzy-graded method based on local texture complexity for adaptive pixel selection, achieving improved imperceptibility at moderate payloads; however, the absence of explicit edge prioritization leads to performance degradation at higher embedding rates. Further, Daiyrbayeva et al. [18] introduced an adaptive reversible data-hiding method for medical X-ray images using region segmentation and interpolation-based embedding, achieving high PSNR and reversibility, but at the cost of increased computational complexity and limited scalability for large-resolution images.

Edge-adaptive strategies have also been enhanced using hybrid detectors and learning-based models [19]. Several studies have employed Sobel, Canny, and Laplacian-based edge guidance to improve imperceptibility [15], [20], but reliance on single or static edge maps limits robustness across diverse image characteristics. The work [21] combined LoG edge detection with pixel-value differencing in RGB images to enhance capacity, but the method is irreversible, restricted to color images, and achieves acceptable quality only at moderate PSNR levels.

To improve embedding efficiency, optimization-based steganography has been widely explored. Singh and Bedi [22] applied particle swarm optimization (PSO) to select embedding pixels, achieving improved imperceptibility, but at the expense of high computational cost and lack of edge prioritization. The authors in [23] employed genetic algorithms for adaptive embedding; however, convergence instability and limited robustness evaluation restrict practical deployment. Omar et al.in [24] combined LoG edge detection with Shark Smell Optimization (SSO) to select embedding pixels, but reliance on LSB substitution and block-wise processing increases overhead and limits scalability under high payload conditions.

To enhance confidentiality, several studies have integrated cryptography with steganography [25], [26]. Some of the studies emphasized encryption as a key security layer in steganographic systems, while Kumar et al. [27] combined AES encryption with adaptive spatial-domain embedding to improve secrecy. Nevertheless, these approaches largely rely on conventional LSB-based embedding, which constrains robustness and payload scalability. Hybrid security systems have also been proposed, such as the DWT-AES-LSB framework by Awadh et al. [28], which improves security and capacity but remains limited by fixed LSB allocation. Survey studies, including Varghese and Sasikala [29], further highlight that existing methods often treat capacity, imperceptibility, optimization, and security as independent objectives rather than as a unified design problem.

The review reveals that LSB methods lack robustness at high payloads, edge-based schemes have limited adaptivity, optimization approaches often ignore encryption, and secure frameworks rely on conventional embedding. Therefore, a unified framework integrating adaptive edge detection, optimization-driven embedding, and strong cryptographic security is required, motivating this work.

III. Proposed Methodology

This section presents the proposed high-capacity edge-adaptive steganography framework with multilayer security. As shown in Fig. 1, it includes preprocessing, edge detection, optimized pixel selection, encryption, embedding, and extraction stages.

Let the grayscale converted cover image, and the secret message can be denoted as Eq. (1) and Eq. (2), respectively.

$$I \in \mathbb{Z}^{M\times N}, I(i,j) \in [0,255] \quad (1)$$

$$\mathcal{M} = \{m_k\}_{k=1}^{L}, m_k \in \{0,1\} \quad (2)$$

The objective is to embed $\mathcal{M}$ into I to generate a stego image I_s, that permits maximum payload capacity, minimum visual distortion, and the payload confidentiality remains preserved. This is formulated as a constrained optimization problem such as [see Eq. (3)]:

$$\max_{\Omega} \; C(\Omega) \text{s.t.} D(I, I_s) \leq \tau \tag{3}$$

where,

- Ω = total embedding pixel set
- $C(\cdot)$ = payload capacity
- $D(\cdot)$ = distortion metric
- τ = imperceptibility threshold

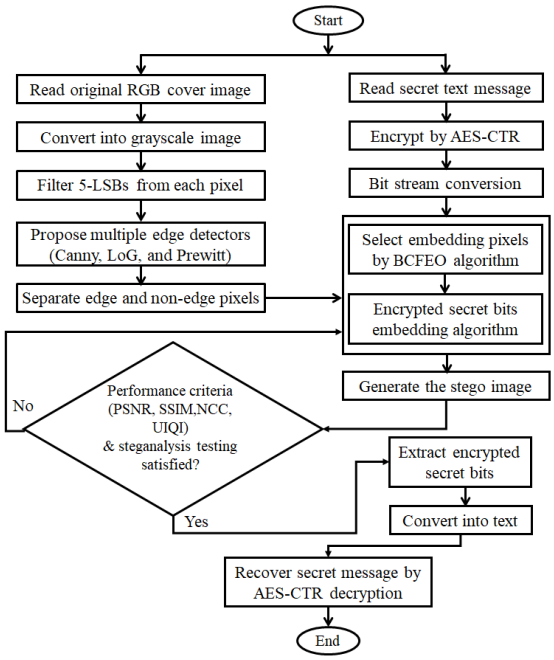


Fig. 1. Flowchart of the proposed edge-adaptive image steganography scheme with BCFEO-based optimized embedding and AES-CTR multilayer security.

A. Cover Image Preprocessing

The RGB cover image is first converted to grayscale to reduce computational complexity and ensure uniform pixel-level processing. A 5-LSB filtering operation is then applied by clearing the five least significant bits of each pixel, as expressed in Eq. (4):

$$I_n(i,j) = \left\lfloor \frac{I(i,j)}{2^5} \right\rfloor \times 2^5 \tag{4}$$

This operation quantizes each 8-bit pixel into multiples of 32, preserving the three most significant bits while discarding low-significance intensity variations. Therefore, the maximum preprocessing-induced pixel deviation is limited to 31 gray levels, while the normalized image remains structurally suitable for stable edge detection and controlled embedding. The purpose of this step is not to preserve exact pixel intensity, but to create a stable reference image in both embedding and extraction phases so that edge-map generation and bit recovery remain consistent. Its effect on image fidelity is subsequently reflected through PSNR, MSE, SSIM, and other quality metrics under different payload capacities. As payload increases, this preprocessing helps limit uncontrolled pixel fluctuation and improves robustness by reducing sensitivity to minor intensity variations.

B. Multi-Detector Edge Identification

Accurate identification of embedding regions is essential for high-capacity steganography. The proposed method employs Canny, LoG, and Prewitt detectors [15], whose outputs are fused to generate a reliable edge map for classifying pixels into edge and non-edge regions. This multi-detector approach improves robustness and adaptability across varying image textures.

1) Canny detector: The Canny method computes image gradients using Sobel operators [30], as in Eq. (5):

$$G = \sqrt{G_x^2 + G_y^2}, \quad \theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{5}$$

where, G_x and G_y are horizontal and vertical gradients. Non-maximum suppression and hysteresis thresholding are then used to retain only strong, well-localized edges. Canny detector ensures strong edge localization and effective noise suppression through gradient analysis and hysteresis thresholding.

2) Laplacian of Gaussian (LoG) detector: The LoG highlights regions of rapid intensity change by convolving the image with [31], see Eq. (6):

$$\nabla^2 G(x,y) = \frac{\partial^2 G}{\partial x^2} + \quad \theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{6}$$

where, $G(x,y)$ is a Gaussian-smoothed image. Zero-crossings of the result indicate edges, allowing detection of fine transitions and weak boundaries. LoG detector highlights fine intensity transitions and captures weak edges often missed by gradient-only detectors.

3) Prewitt detector: The Prewitt operator estimates gradients using convolution masks [3], as in Eq. (7):

$$G_x = I * M_x, \quad G_y = I * M_y \tag{7}$$

with M_x and M_y as horizontal and vertical masks. This operator is computationally simple and effective for detecting directional edges. Prewitt detector computes directional gradients, providing complementary information on vertical and horizontal edge structures.

C. Hybrid Edge Detection

Multiple detectors are used to leverage their complementary strengths: Canny is noise-robust, LoG enhances weak edges but may introduce false responses, and Prewitt captures gradients but is less precise in noise. Their edge maps are computed and combined as follows [see Eq. (8)]:

$$\begin{aligned} E_c &= \text{Canny}(I_n); E_l \\ &= \text{LoG}(I_n); E_p = \text{Prewitt}(I_n) \end{aligned} \tag{8}$$

By fusing their outputs with an OR operation, the system maximizes true edge coverage while minimizing the loss of embedding opportunities that would occur if a single detector were used. This hybridization ensures a richer set of candidate

pixels for embedding, leading to higher payload capacity and greater resilience without compromising imperceptibility. A fused edge map is then obtained as Eq. (9):

$$E_f(i,j) = \begin{cases} 1, & E_c + E_l + E_p \geq \theta \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Pixels are classified as:

- Edge pixels: $\mathcal{E} = \{(i,j) \mid E_f(i,j) = 1\}$
- Non-edge pixels: $\mathcal{N} = \Omega \setminus \mathcal{E}$

D. Optimized Embedding Pixel Selection Using BCFEO

To efficiently utilize the available embedding space, the proposed framework integrates the Bee Colony Footprint Edge Optimization (BCFEO) algorithm for adaptive pixel selection. BCFEO identifies suitable embedding locations by evaluating candidate edge pixels according to their local edge strength, texture variation, and distortion tolerance. Each candidate pixel $p_k \in \mathcal{E}$ is assigned a fitness score, after which the scores are normalized and ranked to categorize pixels based on their embedding suitability. Unlike iterative swarm-position update methods, the proposed BCFEO stage does not physically move pixels; instead, it performs fitness-based categorization of cover-image pixels. The optimized embedding set Ωe^* is selected from the highest-ranked edge pixels using an adaptive threshold determined by the payload requirement. If the capacity of Ωe^* is insufficient, the framework sequentially utilizes the remaining edge pixels Ω_r and, only when necessary, the non-edge pixels Ω_n. This hierarchical selection strategy enables adaptive payload distribution while preserving visual quality by prioritizing pixels with higher distortion tolerance. BCFEO-based pixel score for each candidate pixel $p_k \in \mathcal{E}$ is assigned a fitness score [see Eq. (10)]:

$$F(p_k) = \alpha \cdot | \nabla I_n(p_k) | - \beta \cdot \sigma^2(p_k) \quad (10)$$

where,

- $| \nabla I_n |$ = local gradient magnitude
- σ^2 = local variance
- α, β = weighting factors

The BCFEO iteratively updates pixel selection to maximize the fitness score as Eq. (11):

$$F(p_k)_{max} = \max \sum_{p_k \in \Omega} F(p_k) \quad (11)$$

E. Multilayer Security: AES-CTR Encryption

To enhance confidentiality, the secret message is encrypted prior to embedding using the Advanced Encryption Standard in Counter (AES-CTR) mode. AES-CTR is selected due to its strong cryptographic security, stream-like encryption capability, and absence of padding-induced length expansion that does not impact payload efficiency. The encrypted bitstream exhibits high randomness, preventing meaningful interpretation even if partial extraction occurs. This cryptographic layer forms the first security barrier, complementing the concealment provided by steganographic embedding. The plaintext message $\mathcal{M}$ is encrypted as Eq. (12):

$$C_k = m_k \oplus \text{AES}_K(\text{CTR} + k) \quad (12)$$

where,

- K= secret key
- CTR = initial counter
- C_k= ciphertext bit

CTR mode preserves message length and prevents error propagation.

F. Secret Data Embedding

The encrypted data is converted into a binary bitstream and embedded in the red channel using an adaptive strategy based on pixel classification and optimization output. This allows high payload capacity while preserving imperceptibility and statistical security. The distortion-aware embedding modifies each pixel accordingly, as in Eq. (13):

$$I_s(p) = \begin{cases} I_n(p) + b, & \text{if } b \in \{0,1\} \\ I_n(p), & \text{otherwise} \end{cases} \quad (13)$$

where, b is the encrypted bit. Bit allocation per pixel is adaptive and is determined as [see Eq. (14)]:

$$\eta(p) = \begin{cases} 2, & p \in \Omega_e^* \\ 1, & p \in \Omega_r \\ 0, & p \in \Omega_n \end{cases} \quad (14)$$

The well-defined algorithm for secret data embedding is presented in Algorithm 1.

Algorithm 1: Optimized Edge-Adaptive Embedding with Multilayer Security

Input: Cover image I, secret message $\mathcal{M}$, secret key K, initial counter CTR_0, embedding parameters (edge detectors, BCFEO parameters, payload length L)

Output: Stego image I_s

1: Input cover image I of size $M \times N$

2: Convert I to grayscale image Ig

3: Apply 5-LSB normalization to Ig:

$$I_n(i,j) = floor(\frac{I_g(i,j)}{2^5} 2^5)$$

4: Compute edge maps:

$$E_c \leftarrow Canny(I_n)$$

$$E_l \leftarrow LoG(I_n)$$

$$E_p \leftarrow Prewitt(I_n)$$

5: Fuse edge maps to generate consolidated edge map E_f:

$$E_f(i,j) = 1 \; if \; Ec(i,j) + El(i,j) + Ep(i,j) \geq \theta$$

$$0 \; otherwise$$

6: Classify pixels:

$$Edge \; pixel \; set \; \mathcal{E} \leftarrow \{(i,j) | E_f(i,j) = 1\}$$

$$Non-edge \; pixel \; set \; \mathcal{N} \leftarrow \Omega \setminus \mathcal{E}$$

7: Initialize Bee Colony Footprint Edge Optimization (BCFEO)

8: For each candidate edge pixel $pk \in \mathcal{E}$, compute its local fitness score $F(pk)$ using:

$$F(pk) = \alpha|\nabla In(pk)| - \beta\sigma^2(pk)$$

9: Perform BCFEO-based pixel fitness categorization:

For each candidate edge pixel $pk \in \mathcal{E}$, compute its fitness score $F(pk)$ based on edge strength, local texture variation, and embedding suitability.

Normalize all fitness scores and rank candidate pixels in descending order.

Select optimized embedding pixels as:

$$\Omega e^* = \{ pk \in \mathcal{E} | F_{norm}(pk) \geq \tau \}$$

where τ is an adaptive threshold determined by the required payload length L.

If $|\Omega e^*| < L$, include the next highest-ranked pixels until the payload requirement is satisfied.

10: Define pixels categories as:

$\Omega^* = \Omega e^* \cup \Omega r \cup \Omega n$; where

Ωe^* : BCFEO-optimized edge pixels

Ωr: remaining edge pixels

Ωn: selected non-edge pixels

11: Convert secret message $\mathcal{M}$ into binary bitstream $\{mk\}, k = 1 \dots L$

12: Encrypt bitstream using AES-CTR:

$$ck = mk \oplus AES_K(CTR_0 + k)$$

13: Initialize bit index $k = 1$

14: **for** each pixel $pk \in \Omega^*$ **do**

if $k > L$ **then**

break

end if

15: Determine embedding capacity $\eta(pk)$:

$$\eta(pk) = 2\ bits\ if\ pk \in \Omega e^*$$
$$1\ bit\ if\ pk \in \Omega r$$
$$0\ bit\ if\ pk \in \Omega n$$

16: Embed $\eta(pk)$ encrypted bits into pk in red channel of I :

$$I'_R(pk) \leftarrow I_R(pk) + \Sigma\ (ck \times 2^b), b \in \{0, \dots, \eta(pk) - 1\}$$

17: $Update\ k \leftarrow k + \eta(pk)$

18: end for

19: Reconstruct stego image I_s through merging I'_R and remaining two unchanged channels

20: Output stego image I_s

G. Secret Data Extraction and Message Recovery

At the receiver, extraction follows the inverse process. Using synchronized parameters, embedded bits are retrieved from the red channel and decrypted via AES-CTR to recover the original message. The extraction and recovery are given in Eq. (15) and Eq. (16):

$$\hat{b}_k = I_s(p_k) mod 2 \quad (15)$$

$$\hat{m}_k = \hat{b}_k \oplus \text{AES}_K(\text{CTR} + k) \quad (16)$$

The secret data extraction and reliable message recovery algorithm with a well-structured definition is presented in Algorithm 2.

Algorithm 2: Edge-Adaptive Extraction and AES-CTR Decryption

Input: Stego image I_s, secret key K, initial counter CTR_0, embedding parameters (edge detectors, BCFEO parameters, payload length L)

Output: Recovered secret message $\hat{\mathcal{M}}$

1: Convert I_s to grayscale image I_sg

2: Apply 5-LSB clearing to $I_sg \rightarrow I_sn$

3: Compute edge maps:

$$E_c \leftarrow Canny(I_sn)$$
$$E_l \leftarrow LoG(I_sn)$$
$$E_p \leftarrow Prewitt(I_sn)$$

4: Fuse edge maps to generate E_f:

$$E_f(i,j) = 1\ if\ Ec(i,j) + El(i,j) + Ep(i,j) \geq \theta$$
$$0\ otherwise$$

5: Classify pixels:

$$Edge\ pixel\ set\ \mathcal{E} \leftarrow \{ (i,j) | E_f(i,j) = 1 \}$$
$$Non-edge\ pixel\ set\ \mathcal{N} \leftarrow \Omega \setminus \mathcal{E}$$

6: Initialize Bee Colony Footprint Edge Optimization (BCFEO)

7: For each candidate edge pixel $pk \in \mathcal{E}$, compute fitness:

$$F(pk) = \alpha|\nabla In(pk)| - \beta\sigma^2(pk)$$

8: Perform BCFEO-based pixel fitness categorization:

$\Omega^* = \Omega e^* \cup \Omega r \cup \Omega n$; where

Ωe^* : BCFEO-optimized edge pixels

Ωr: remaining edge pixels

Ωn: selected non-edge pixels

9: Initialize empty ciphertext bitstream $\hat{\mathcal{C}} = \emptyset$

10: Select the modified red channel I'_R of stego image

11: **for** $k = 1$ to L **do**

12: Select extraction pixel $pk \in \Omega^*$

Extract encrypted bit:

13: $\hat{c}k = I_s(pk) \bmod 2$

14: Append $\hat{c}k$ to $\hat{C}$

15: end for

16: Group $\hat{C}$ into byte-aligned blocks

17: Initialize AES-CTR keystream using key K and counter CTR_0

18: **for** each encrypted bit $\hat{c}k$ **do**

Generate keystream bit:

$$zk = AES_K(CTR_0 + k)$$

Recover plaintext bit:

$$\hat{m}k = \hat{c}k \oplus zk$$

end for

19: Reconstruct original message $\hat{\mathcal{M}}$ from bitstream $\{\hat{m}k\}$

20: Output recovered message $\hat{\mathcal{M}}$

IV. Results and Analysis

The proposed framework was implemented and evaluated in Python 3.11.7 using the Anaconda distribution. Experiments were conducted on standard RGB benchmark images obtained from the USC-SIPI image database [32], together with additional natural images captured independently to introduce scene diversity beyond benchmark datasets. For consistent comparative evaluation, all color images were resized to 256×256 and converted into 8-bit grayscale images, resulting in 65,536 pixels per image for each simulation. Among the complete set of tested images, six representative samples, namely "Pepper", "F16", "Sailboat", "Boats", and two personally captured natural images, "KL Tower" and "UKM", are reported in detail due to space limitations and presentation clarity. These images were selected to cover different structural and textural characteristics, including smooth regions, strong edges, object boundaries, and complex natural backgrounds. Nevertheless, the evaluation is primarily based on fixed-resolution grayscale images; therefore, future work will extend the analysis to larger image sets with different resolutions, content categories, and statistical properties to further validate the generalizability of the proposed framework.

A. Payload Capacity Analysis

The achievable payload capacity of an image steganography system is primarily governed by the number and spatial distribution of pixels deemed suitable for embedding. In the proposed framework, payload capacity is directly influenced by the effectiveness of edge and non-edge pixel identification, which is performed exclusively on a reference intensity plane prior to data embedding. To evaluate the baseline embedding potential, edge pixels were initially identified using individual edge detectors, namely Canny, Prewitt, and Laplacian of Gaussian (LoG) and their hybridization applied to consecutive processed cover images.

TABLE I. Edge Pixels Found by Different Detectors and their Hybridization in Original and Cleared Images.

Images (65536 pixels)	Canny	LoG	Prewitt	Hybrid fused	
	Original grayscale (Org)			Org	Cleared
Pepper	7193	9284	8042	14132	21498
F16	6400	10607	9985	14246	16554
Sailboat	10774	12112	12666	20388	25362
Boats	9425	13211	11498	19515	23565
KL_tower	7151	11075	11239	17653	22798
UKM	13307	13295	12849	23442	27086

The results summarized in Table I indicate that the number of detected edge pixels varies significantly across detectors and image content. The corresponding edge pixel counts obtained from these representations are illustrated in Fig. 2, where the hybrid approach on the cleared image consistently yields a higher and more stable number of embedding candidates across all benchmark images, leading to the higher embedding capacity achieved in the adaptive mode of embedding for the proposed system.

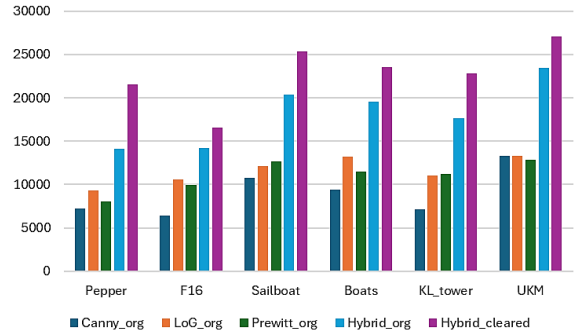


Fig. 2. Visual presentation of edge pixels found in different detectors and their hybrid fusion.

B. Imperceptibility (Qualitative) Analysis

As an initial qualitative assessment of imperceptibility, the visual quality of the generated stego images is examined from the perspective of the Human Visual System (HVS) before proceeding to quantitative performance evaluation. Fig. 3 illustrates representative examples of four original cover images and their corresponding stego images produced at an embedding payload of 1 bpp.

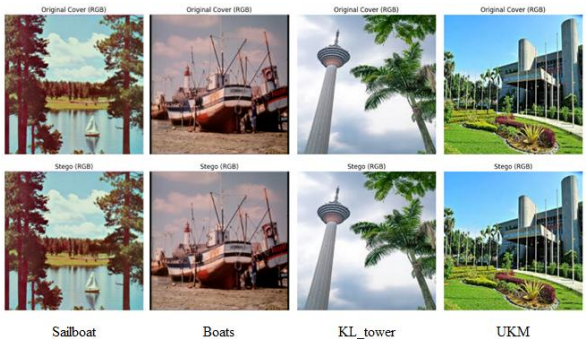


Fig. 3. Visual imperceptibility comparison between original cover images (top row) and corresponding stego images (bottom row) generated using the proposed steganography framework. Test images include Sailboat, Boats, KL_tower, and UKM.

A careful visual comparison between the cover and stego images reveals no discernible visual degradation across all evaluated test images. In particular, no visible noise patterns, color shifts, edge blurring, blocking artifacts, or contrast inconsistencies are observed in the stego images. This observation holds consistently for images with varying visual characteristics, including smooth regions, textured areas, object boundaries, natural landscapes, and human facial features. From an HVS standpoint, the stego images remain visually indistinguishable from their original counterparts.

C. Imperceptibility (Quantitative) Analysis

Imperceptibility is a critical requirement in image steganography, ensuring that the embedded payload does not introduce perceptible visual or statistical artifacts. Table II reports the detailed imperceptibility performance of the proposed method under varying payload sizes using the F-16 benchmark image, while Table III provides a cross-image summary of PSNR and SSIM values for six representative test images at selected payload capacities.

TABLE II. PRESENTATION OF PSNR, SSIM, AND NCC VALUES UNDER DIFFERENT PAYLOAD REGIMES IN F-16 IMAGE.

Payload Sizes/rates	PSNR	SSIM	NCC	UIQI	Entropy deviation
128 bits	83.91	1.0000	1.0000	1.0000	0.0000
256 bits	80.24	1.0000	1.0000	1.0000	0.0000
512 bits	77.00	1.0000	1.0000	1.0000	0.0000
1024 bits	74.07	1.0000	1.0000	1.0000	0.0000
0.031 bpp	70.91	1.0000	1.0000	1.0000	0.0000
0.063 bpp	67.81	1.0000	1.0000	1.0000	0.0000
0.125 bpp	64.96	0.9999	1.0000	1.0000	0.0000
0.25 bpp	62.01	0.9998	1.0000	1.0000	0.0001
0.50 bpp	58.92	0.9996	1.0000	1.0000	0.0001
1.00 bpp	55.91	0.9989	1.0000	1.0000	0.0008
2.00 bpp	48.90	0.9947	0.9999	0.9999	0.0012
3.00 bpp	42.61	0.9795	0.9991	0.9997	0.0141
4.00 bpp	36.67	0.9406	0.9963	0.9991	0.0333

As shown in Table II, the PSNR values for the F-16 image remain exceptionally high at low payloads, exceeding 80 dB for 128-bit embedding and gradually decreasing with increasing payload. Even at moderate payloads of 0.25-0.50 bpp, the PSNR values remain above 58 dB, which is well beyond the commonly accepted imperceptibility threshold of 40 dB. At higher payloads of 1-4 bpp, PSNR decreases as expected due to increased embedding density; however, the values remain within acceptable limits, demonstrating controlled distortion even under aggressive embedding conditions.

The SSIM values also remain extremely close to unity across most payload regimes, with only marginal degradation at higher payloads. This indicates that the proposed edge-adaptive and optimized embedding strategy effectively preserves the structural and perceptual characteristics of the cover image. Similarly, the NCC and UIQI values remain near 1.0 for all payloads, confirming strong statistical similarity between the cover and stego images. The entropy deviation remains negligible at low and moderate payloads and increases only slightly at very high embedding rates, indicating minimal disturbance to the underlying pixel distribution.

TABLE III. PRESENTATION OF PSNR AND SSIM VALUES UNDER DIFFERENT PAYLOAD REGIMES IN DIFFERENT TEST IMAGES.

Images	PSNR and SSIM for different payload rates					
	0.125 bpp		1.00 bpp		3.00 bpp	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Pepper	64.89	0.9999	55.90	0.9992	42.61	0.9761
F16	64.96	0.9999	55.91	0.9989	42.66	0.9795
Sailboat	64.97	0.9999	55.93	0.9992	42.61	0.9848
Boats	64.93	0.9999	55.93	0.9991	42.65	0.9862
KL_tower	64.97	0.9999	55.89	0.9989	42.63	0.9706
UKM	65.02	0.9999	55.91	0.9993	42.61	0.9796
Average	64.96	0.9999	55.91	0.9992	42.61	0.9795

To further address cross-image consistency, Table III summarizes PSNR and SSIM results across Pepper, F16, Sailboat, Boats, KL_tower, and UKM images at 0.125, 1.00, and 3.00 bpp. The average PSNR values are 64.96 dB, 55.91 dB, and 42.61 dB, respectively, while the corresponding average SSIM values are 0.9999, 0.9992, and 0.9795. These aggregated results confirm that the proposed method maintains high imperceptibility across different image contents, including benchmark images and independently captured natural images. The small variation among the tested images further indicates that the method is not strongly dependent on a single image structure or texture pattern.

The combined results in Table II and Table III demonstrate that the proposed method achieves a favorable capacity-imperceptibility balance across a wide range of payload capacities. The F-16 and cross-image results confirm that the proposed multi-detector edge selection and BCFEO-optimized embedding strategy maintain consistent image quality under increasing payloads across different test images.

D. Histogram Consistency Analysis

Histogram-based steganalysis evaluates whether embedding introduces noticeable changes in pixel intensity distributions. As shown in Fig. 4, the RGB histograms of the stego images closely match those of the corresponding cover images across all channels. The peak positions, intensity spread, and overall distribution shapes remain consistent, with no visible peak distortion, flattening, or abnormal clustering. This strong alignment indicates that the proposed method preserves first-order statistical properties and does not introduce detectable artifacts, even under multi-channel embedding.

Further strengthening the histogram-based security assessment, signed Pixel Difference Histogram (PDH) plots are analyzed at the same embedding rate and presented in Fig. 5. It demonstrates a sharp, symmetric peak centered at zero for both cover and stego images, indicating that most neighboring pixel differences remain small and evenly distributed around zero. The near-perfect overlap of the signed PDH curves confirms that

the embedding process does not introduce directional bias or systematic local distortions. The combined evidence from RGB histogram comparison and PDH analysis demonstrates that the proposed steganographic framework effectively resists histogram-based steganalysis.

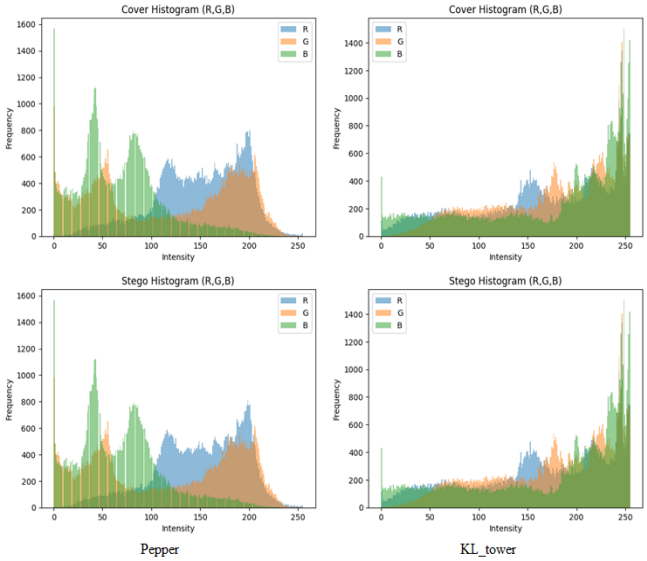


Fig. 4. RGB histogram comparison of cover (top row) and corresponding stego images (bottom row) for Pepper and KL_tower images, showing minimal statistical deviation and preservation of intensity distributions after embedding.

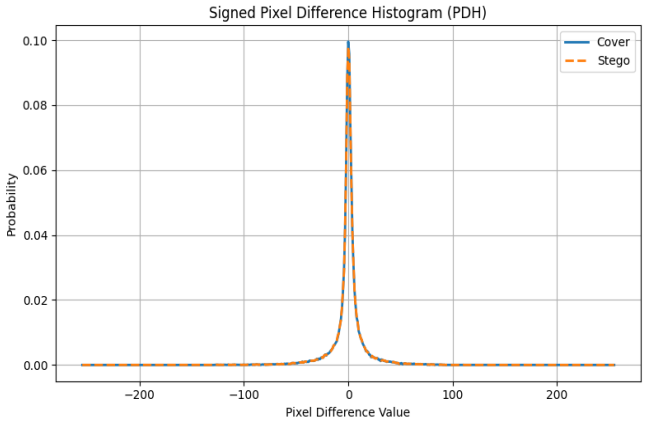


Fig. 5. PDH analysis of the cover and the stego image (F-16) at a data hiding rate of 1.00 bpp.

E. Regular-Singular (R-S) Steganalysis Evaluation

Regular-Singular (R-S) steganalysis assesses higher-order statistical changes in pixel groups under flipping masks. As shown in Fig. 6, the proportions of regular (R) and singular (S) groups for both masks $M = [1, 0, 1, 0]$ and $M = [0, 1, 0, 1]$, remain closely aligned between cover and stego images, with minimal and symmetric differences. Specifically, the relative differences between R_M and S_M, as well as between R_{-M} and S_{-M}, are minimal and symmetric across the cover-stego pairs. Additionally, the proportion of unusable groups (U) remains zero in all cases, indicating that the embedding process preserves natural image statistics without introducing detectable structural distortions.

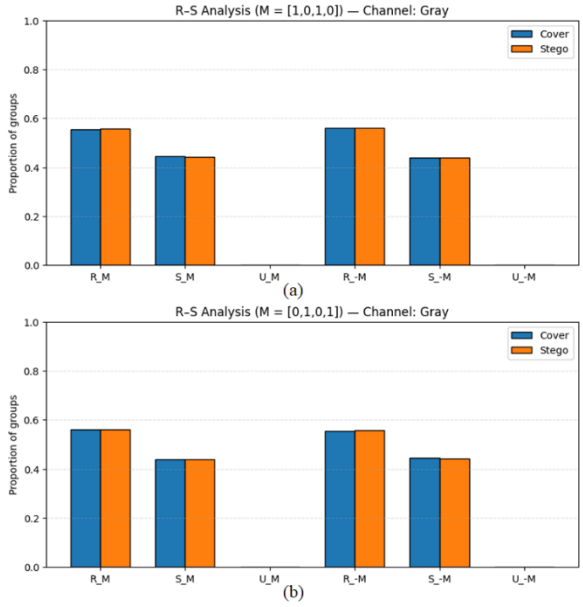


Fig. 6. Graphical presentation of Regular-Singular (R-S) steganalysis results for the grayscale converted image at an embedding rate of 1 bpp using two complementary flipping masks: (a) M=[1,0,1,0] and (b) M=[0,1,0,1].

The close correspondence between the proportions of regular (R), singular (S), and unusable (U) groups for the cover and stego images demonstrate preservation of structural statistics and resistance to R-S based steganalysis. The R-S analysis shows that the proposed embedding strategy preserves the natural imbalance between regular and singular groups across multiple flipping masks, thereby avoiding the characteristic equalization behavior exploited by R-S steganalysis.

F. Chi-Square and Pairs-of-Values (PoV) Steganalysis Evaluation

Chi-square (χ^2) steganalysis was conducted on the complete image set; representative results are shown for clarity and space efficiency. The global chi-square statistics for the grayscale converted "Sailboat" cover and stego image at 1.00 bpp embedding payload are derived from the grayscale intensity distribution after excluding low-frequency pairs is selected in an experimental simulation.

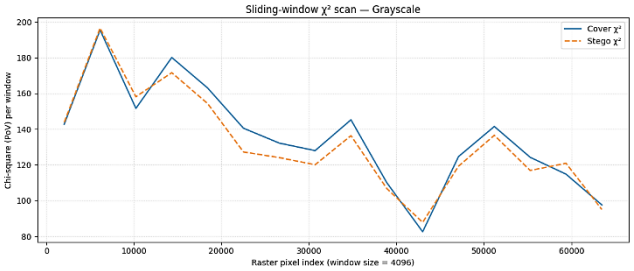


Fig. 7. Sliding-window (size=4096) chi-square ($\chi 2$) scan of the grayscale (Sailboat) cover and stego images at an embedding rate of 1 bpp.

To assess localized detectability, a sliding-window chi-square scan is performed across the rasterized grayscale image and illustrated in the sliding-window χ^2 plot shown in Fig. 7. The resulting chi-square trajectories of the cover and stego

images with 65536 hidden bits (1.00 bpp) exhibit closely aligned fluctuation patterns over all window positions and no persistent divergence or abrupt spike is observed in the stego curve relative to the cover curve. The close alignment of the chi-square trajectories indicates the absence of localized statistical anomalies introduced by the embedding process.

G. Crypto-Security and Reliability Analysis

Instead of direct plaintext embedding, the proposed system employs AES in Counter (CTR) mode to encrypt the secret message prior to bitstream generation. This ensures confidentiality while remaining independent of the embedding process. Importantly, AES-CTR preserves exact length equivalence between plaintext and ciphertext. For example, a 256-character message produces a 2048-bit stream, and the encrypted output retains the same 2048-bit length after binary conversion. As illustrated in Fig. 8, this one-to-one correspondence confirms strict length preservation throughout the process.

The results confirm that the recovered plaintext is identical to the original message, validating the reliability of the complete pipeline, including encryption, embedding, extraction, and decryption. As illustrated in Fig. 8, a 256-character message undergoes encryption, embedding, and recovery, with perfect agreement between the original and extracted plaintext, ensuring exact semantic reconstruction.

This accuracy is achieved while maintaining strict length consistency throughout all stages. The integration of AES-CTR encryption enhances security by protecting confidentiality, while its length-preserving property ensures that embedding capacity is unaffected. Consequently, the proposed framework achieves a balanced trade-off between security, capacity, and reliable recovery, reinforcing its effectiveness for secure communication.



Fig. 8. Authorized receiver validation showing end-to-end exact message (256 character) recovery, including plaintext encryption, ciphertext-to-bitstream embedding, extraction, and AES-CTR decryption.

H. Comparison with Existing Steganographic Methods

Table IV compares the proposed method with representative state-of-the-art image steganography approaches in terms of embedding capacity and imperceptibility. To ensure transparency, the comparison is interpreted in two ways based on the relative capacity-imperceptibility performances. The proposed method was evaluated under a consistent experimental setting, and the comparing values are taken from their original publications as several works do not provide complete implementation details, source codes, or identical parameter settings. Therefore, the comparison is intended to show the relative capacity-imperceptibility trend rather than a fully controlled one-to-one reproduction of all methods.

Existing hybrid edge-based methods achieve high visual quality, with PSNR values around 60.6 dB, but they operate at very low payloads, typically ≤ 0.125 bpp, which limits their applicability for high-capacity data hiding. In contrast, RGB model-based edge detection methods support higher payloads of up to 2.80 bpp but experience noticeable visual degradation, as indicated by lower PSNR and SSIM values. The IWT-PSO method provides competitive imperceptibility at low to moderate payloads, particularly within the 0.25-0.75 bpp range, where PSNR remains above 50 dB. However, its performance decreases as payload increases, and it is not extensively reported for higher-capacity embedding.

TABLE IV. COMPARISON OF THE PROPOSED METHOD TO OTHER COMPETING METHODS.

Method	Embedding rate (bpp)	PSNR (dB)	SSIM
Hybrid edge detection [16]	0.125	60.58	0.9992
RGB model-based edge detection [21]	2.80	33.51	0.8600
IWT-PSO [8]	0.25	56.53	0.9999
	0.61	52.78	0.9999
	0.75	50.68	0.9999
Proposed	0.125	64.96	0.9999
	0.25	62.01	0.9998
	0.50	58.92	0.9996
	1.00	55.91	0.9989
	2.00	48.90	0.9947
	3.00	42.61	0.9795
	4.00	36.67	0.9406

By comparison, the proposed method maintains a favorable capacity-imperceptibility balance across a wider payload range. At low payloads, it achieves high PSNR and near-unity SSIM, while at higher payloads of 1-4 bpp, it continues to preserve acceptable visual quality where many existing methods either degrade substantially or do not report results. This improvement is mainly attributed to the integration of multi-detector edge adaptivity and BCFEO-based optimized pixel selection, which prioritizes pixels with stronger edge characteristics and higher distortion tolerance. The results indicate that the proposed framework provides competitive and scalable capacity-imperceptibility trade-off for high-capacity steganographic embedding.

V. Conclusion

This study proposed a multilayer secure image steganography framework based on edge-adaptive embedding and pre-encryption for high-capacity covert communication. The framework combines hybrid edge detection using Canny, LoG, and Prewitt operators with Bee Colony Footprint Edge Optimization (BCFEO) to identify and prioritize embedding locations with high distortion tolerance. In addition, AES-CTR pre-encryption was incorporated to strengthen payload confidentiality while preserving exact message length and enabling reliable recovery. Experimental results demonstrated that the proposed method achieved high embedding capacity with controlled visual distortion, high structural similarity, and strong resistance to classical statistical steganalysis. Comparative evaluation further showed that the method provides a favorable capacity-imperceptibility trade-off over representative existing approaches, particularly at higher payloads. However, higher payloads gradually reduce imperceptibility, as indicated by lower PSNR and SSIM. The integration of adaptive edge selection, optimization-driven embedding, and length-preserving cryptographic protection makes the proposed framework an effective solution for secure image-based communication. The future work will validate the method on larger, higher-resolution, and more diverse datasets, evaluating robustness against different threats of adversarial or learning-based detection models.

Acknowledgment

This research was funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT01/UKM/02/4), with research facilities provided by Universiti Kebangsaan Malaysia (UKM). The author, A. F. M. Zainul Abadin, gratefully acknowledges the ICT Division, the Ministry of Posts, Telecommunications and Information Technology, Bangladesh, for awarding him a PhD fellowship (FY 2022-2023, third round).

References

[1] S. Wang, J. Wang, and Z. Yu, "Privacy-preserving authentication in wireless IoT: applications, approaches, and challenges," IEEE Wirel. Commun., vol. 25, no. 6, pp. 60–67, 2019.

[2] J. Chen et al., "Fused Fuzzy Deep Learning and Information Steganography for Privacy Preservation in Medical Consumer Electronics," IEEE Trans. Consum. Electron., vol. 71, no. 2, pp. 7105–7116, 2025, doi: 10.1109/TCE.2025.3525536.

[3] A. F. M. Z. Abadin, R. Sulaiman, and M. K. Hasan, "A High-Security Image Steganography System Employing Multiple Edge Detectors," in Communications in Computer and Information Science, CCIS, vol. 2597, Springer Nature Singapore, 2026, pp. 40–55. doi: 10.1007/978-981-95-0172-4_3.

[4] M. Njoum, R. Sulaiman, Z. Shukur, and F. Qamar, "High-Secured Image LSB Steganography Using AVL-Tree with Random RGB Channel Substitution," Comput. Mater. Contin., vol. 81, no. 1, pp. 183–211, 2024, doi: 10.32604/cmc.2024.050090.

[5] A. H. Mohsin et al., "New Method of Image Steganography Based on Particle Swarm Optimization Algorithm in Spatial Domain for High Embedding Capacity," IEEE Access, vol. 7, pp. 168994–169010, 2019, doi: 10.1109/ACCESS.2019.2949622.

[6] W. Luo, F. Huang, and J. Huang, "Edge adaptive image steganography based on lsb matching revisited," IEEE Trans. Inf. Forensics Secur., vol. 5, no. 2, pp. 201–214, 2010, doi: 10.1109/TIFS.2010.2041812.

[7] X. Duan, D. Guo, N. Liu, B. Li, M. Gou, and C. Qin, "A New High Capacity Image Steganography Method Combined with Image Elliptic Curve Cryptography and Deep Neural Network," IEEE Access, vol. 8, pp. 25777–25788, 2020, doi: 10.1109/ACCESS.2020.2971528.

[8] W. Wang and Q. Li, "An Image Steganography Algorithm Based on PSO and IWT for Underwater Acoustic Communication," IEEE Access, vol. 10, pp. 107376–107385, 2022, doi: 10.1109/ACCESS.2022.3212691.

[9] S. Ghoul, R. Sulaiman, and Z. Shukur, "A Review on Security Techniques in Image Steganography," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 361–385, 2023, doi: 10.14569/IJACSA.2023.0140640.

[10] O. Cetin and A. T. Ozcerit, "A new steganography algorithm based on color histograms for data embedding into raw video streams," Comput. Secur., vol. 28, no. 7, pp. 670–682, 2009, doi: 10.1016/j.cose.2009.04.002.

[11] A. F. M. Z. Abadin, R. Sulaiman, and M. K. Hasan, "Randomization Strategies in Image Steganography Techniques : A Review," Comput. Mater. Contin., vol. 80, no. 2, pp. 3139–3171, 2024, doi: 10.32604/cmc.2024.050834.

[12] C. K. Chan and L. M. Cheng, "Hiding data in images by simple LSB substitution," Pattern Recognit., vol. 37, no. 3, pp. 469–474, 2004, doi: 10.1016/j.patcog.2003.08.007.

[13] S. Rahman et al., "A novel and efficient digital image steganography technique using least significant bit substitution," Sci. Rep., vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-024-83147-3.

[14] A. F. M. Z. Abadin, R. Sulaiman, and M. K. Hasan, "Enhanced Predictive Pixel Deviation Frame Guided Hybrid Edge-Based Image Steganography for High Capacity and Imperceptibility," in Proc. The 10th International Conference on Electrical Engineering and Informatics, ICEEI 2025, Kuching, Malaysia, IEEE, 2025, pp. 1–6. doi: 10.1109/ICEEI68459.2025.11330299.

[15] H. Sultana, A. H. M. Kamal, G. Hossain, and M. A. Kabir, "A Novel Hybrid Edge Detection and LBP Code-Based Robust Image Steganography Method," Futur. Internet, vol. 15, no. 3, pp. 1–23, 2023, doi: 10.3390/fi15030108.

[16] M. Habiban, F. R. Hamade, and N. A. Mohsin, "Hybrid Edge Detection Methods in Image Steganography for High Embedding Capacity," Cybern. Inf. Technol., vol. 24, no. 1, pp. 157–170, 2024, doi: 10.2478/cait-2024-0009.

[17] D. R. I. M. Setiadi, S. Rustad, P. N. Andono, and G. F. Shidik, "Graded fuzzy edge detection for imperceptibility optimization of image steganography," Imaging Sci. J., vol. 72, no. 6, pp. 693–705, 2024, doi: 10.1080/13682199.2023.2219880.

[18] E. Daiyrbayeva et al., "An Adaptive Steganographic Method for Reversible Information Embedding in X-Ray Images," Computers, vol. 14, no. 9, pp. 1–23, 2025, doi: 10.3390/computers14090386.

[19] M. Kumar, S. K. Singh, and S. Kim, "Hybrid deep learning-based cyberthreat detection and IoMT data authentication model in smart healthcare," Futur. Gener. Comput. Syst., vol. 166, p. 107711, 2025.

[20] A. F. M. Z. Abadin, M. K. Hasan, and R. Sulaiman, "An Enhanced Payload Image Steganography Employing Hybrid Edge Detection Technique and MSB Cover Image," in Communications in Computer and Information Science, CCIS, vol. 2597, Springer Nature Singapore, 2026, pp. 284–300. doi: 10.1007/978-981-95-0172-4_19.

[21] H. H. Liu and Y. T. Su, "Color Image Steganography Method Based on RGB Model and Edge Detection," Multimed. Tools Appl., no. 0123456789, 2024, doi: 10.1007/s11042-024-20008-1.

[22] S. Kumar, S. Pujari, A. Kulkarni, P. Misal, and I. R. Kale, "Image Steganography Using Particle Swarm Optimization," in International Conference on Information Science and Applications, 2023, pp. 115–130.

[23] S. M. Abdulmaged and N. M. Abdulmaged, "A new steganography technique based on genetic algorithm," Glob. J. Eng. Technol. Adv., vol. 16, no. 2, pp. 135–139, 2023.

[24] P. J. Karim, D. R. Arif, A. O. Abdalrahman, O. Y. Abdulhammed, T. S. Ali, and A. A. Saffer, "A Secure Image Steganography Using Shark Smell Optimization and Edge Detection Technique," Kurdistan J. Appl. Res., vol. 7, no. 2, pp. 11–25, 2022, doi: 10.24017/Science.2022.2.2.

[25] S. Kaur, S. Singh, M. Kaur, and H. N. Lee, "A Systematic Review of Computational Image Steganography Approaches," Arch. Comput. Methods Eng., vol. 29, no. 7, pp. 4775–4797, 2022, doi: 10.1007/s11831-022-09749-0.

[26] M. Njoum, R. Sulaiman, Z. Shukur, and F. Qamar, "A high-capacity image steganography based on random separate hashing data structure with secure CHACHA20-DRBG and AES encryption algorithms," Array, vol. 30, no. April, p. 100863, 2026, doi: 10.1016/j.array.2026.100863.

[27] A. Kumar, P. S. Sekhar, D. Khatua, D. K. Prasad, and A. A. Sekh, "Enhanced Image Steganography Using Secure Random Pixel Distribution," Secur. Priv., vol. 9, no. 1, p. e70186, 2026.

[28] [28] information security system via combination of compression, cryptography, and image steganography," Int. J. Electr. Comput. Eng., vol. 12, no. 6, pp. 6574–6584, 2022, doi: 10.11591/ijece.v12i6.pp6574-6584.

[29] F. Varghese and P. Sasikala, "A Detailed Review Based on Secure Data Transmission Using Cryptography and Steganography," Wirel. Pers. Commun., vol. 129, no. 4, pp. 2291–2318, 2023, doi: 10.1007/s11277-023-10183-z.

[30] Y. Song, C. Li, S. Xiao, Q. Zhou, and H. Xiao, "A parallel Canny edge detection algorithm based on OpenCL acceleration," PLoS One, vol. 19, no. 1, p. e0292345, 2024.

[31] S. K. Ghosal, J. K. Mandal, and R. Sarkar, "High payload image steganography based on Laplacian of Gaussian (LoG) edge detector," Multimed. Tools Appl., vol. 77, no. 23, pp. 30403–30418, 2018.

[32] "The USC-SIPI Image Database," University Southern California Signal and Image Processing Institute. [Online]. Available: https://sipi.usc.edu/database/

Multimodal Machine Learning for Cybersecurity in Internet of Things Environments: A Literature Review

Abdelaaziz NASSIRI, Azeddine Khyat, Kama El Guemmat, Mohamed Aazi
Computing, Artificial Intelligence and Cyber Security Laboratory (2IACS),
ENSET Mohammedia, Hassan II University of Casablanca, Morocco

***Abstract*—This research provides a comprehensive synthesis of Multimodal Machine Learning (MML) as a transformative paradigm for IoT defense. By integrating heterogeneous data streams, including network flow statistics, device-level telemetry, and behavioral biometrics, MML architectures facilitate a holistic understanding of system states. The algorithmic advancements that were analyzed are classified into hybrid CNN-RNN structures and state-of-the-art cross-modal Transformers, and evaluate their performance across benchmark datasets such as ToN-IoT and CICIoT2023. Quantitative results show that cross-modal Transformers achieve F1-scores between 0.95 and 0.99 across detection tasks, while hybrid CNN-LSTM models range from 0.89 to 0.96. Furthermore, this study addresses the technical "optimization triad" of pruning, quantization, and edge-cloud orchestration required to deploy these models on resource-constrained hardware.**

***Keywords*—Internet of Things (IoT); Multimodal Machine Learning; deep learning, cyber-physical security; data fusion; Intrusion Detection Systems (IDS); Edge AI, Explainable AI (XAI)**

I. Introduction

The digital landscape is currently undergoing a fundamental reconfiguration, driven not by incremental expansion but by the convergence of ubiquitous connectivity, exponential data proliferation, and transformative advancements in computational intelligence. Central to this paradigm shift is the Internet of Things (IoT), a technological framework that has evolved from a disparate collection of connected devices into a central component of contemporary digital infrastructure.

As billions of autonomous nodes from tiny environmental sensors to heavy-duty industrial actuators stream data in real-time, the traditional boundary between the digital and the physical is effectively dissolving into a cyber-physical continuum.

However, security frameworks have struggled to keep pace with this rapid expansion. The massive, fragmented attack surface presents challenges that traditional network defense mechanisms cannot adequately address [1] [7].

A. Context: IoT Convergence and Security Bottlenecks

The rapid integration of IoT into critical infrastructure, including power grids, healthcare systems, and transportation networks, has fundamentally altered the risk profile of modern civilization. In these high-stakes environments, security is not merely a matter of data privacy; it constitutes a non-negotiable requirement for operational safety.

The Industrial Internet of Things (IIoT) illustrates this dynamic. Industrial systems are exposed to cyber threats that can result in physical consequences, as the integration of physical machinery with intelligent electronics has yielded significant efficiency improvements. A potential physical catastrophe can result not only from a software defect but also from a digital exploit within a control loop [4] [5] [6] [8].

The IoT ecosystem introduces several structural challenges that confound traditional IT security approaches:

- Extreme Heterogeneity: The wide variety of hardware architectures and proprietary protocols renders the application of a universal security standard practically infeasible.
- Resource Constraints: Many devices operate under severe power and memory limitations, preventing the execution of standard encryption or heavy anti-malware suites.
- Data Volume: With an estimated 26 billion devices expected by 2030, the resulting data volume makes manual oversight impossible. The industry is now forced to rely on automated, intelligent defense systems [2] [3] [9].

B. Problem Statement: The Limitations of Legacy Defense

Traditional cybersecurity, built for the static "walled garden" networks of a previous decade, faces significant challenges in addressing the dynamic complexity of IoT. Most current systems still rely on unimodal analysis that examines a single stream, such as network logs, to detect anomalies. While this approach keeps computational costs low, it creates a blind spot: it cannot detect multi-vector attacks that span different system layers [1] [9].

Despite being the most widely deployed approach, signature-based IDS is not effective against zero-day exploits. Anomaly-based models often cannot distinguish between benign deviations from baseline behavior and actual attacks, leading to an excessive amount of false positive alerts. These systems have a tendency to fail to distinguish between a true intrusion and harmless network latency or transient sensor malfunction [2] [5] [35] [37].

This limitation of unimodal approaches creates vulnerabilities that sophisticated attackers can exploit. An adversary can execute a sequence of actions that appear benign individually such as a minor CPU spike combined with a small shift in packet timing but collectively indicate a security breach. Without the ability to correlate these scattered signals, traditional systems lack the holistic situational awareness needed to defend complex IoT deployments [17] [18] [12].

C. Motivation: Why Multimodal Approaches are the Future?

The limitations of single-source models have motivated a shift toward Multimodal Machine Learning (MML). MML provides a unified framework to ingest and analyze diverse data types simultaneously. By fusing network traffic, device telemetry, and even behavioral biometrics, MML systems can catch the subtle indicators of an attack that are invisible to a unimodal filter [12] [18] [19].

The fundamental principle is data complementarity. Information that appears as noise in one channel can be clarified by another. For example, a suspected DDoS attack on the network can be validated by a corresponding increase in a device's heat profile (high temperature alarm) or a rapid decrease in accessible RAM. This cross-modal validation enhances detection and significantly reduces false positive rates [2] [11].

1) Contribution and Scope: This review integrates the technological frontier at the intersection of MML and IoT security. The focus is exclusively on the mechanics of data fusion rather than doing a comprehensive study; all models were classified from hybrid deep learning to Transformer-based networks, addressing the challenging issue of executing these resource-intensive models on constrained edge devices [3] [7] [9].

2) Organization of the paper: The remainder of this document is organized as follows: Section II establishes the preliminaries, detailing the IoT threat landscape, the variety of data modalities available, and the core strategies of multimodal fusion. Section III presents the systematic review methodology. Section IV provides taxonomy and classification of MML architecture. Section V discusses applications and use cases. Section VI reviews datasets and evaluation metrics. Section VII provides a critical discussion of trade-offs and limitations. Section VIII identifies open challenges and future directions. Section IX concludes the review.

II. Background And Preliminaries

Developing an effective multimodal security framework requires a comprehensive understanding of the operational environment, the nature of the data involved, and the mathematical mechanisms for integrating that data [1] [3] [7].

A. IoT Security Landscape: Major Threats and Vulnerabilities

The IoT threat landscape is characterized by its breadth and the potential consequences of successful exploitation. Threats are typically categorized based on their intent, origin, and the specific layer of the IoT stack they target [1] [3] [22]. Table I presents the IoT threat taxonomy.

TABLE I. IoT Threat Taxonomy

Threat Category	Targeted Layer	Mechanism	Primary Impact
DDoS Attacks	Network	Overwhelming target resources using a botnet.	System unavailability and network collapse.
Malware/ Botnets	Application/Edge	Mirai, BASHLITE, or ransomware infecting devices.	Information leakage and device takeover.
False Data Injection	Physical/ Sensing	Injecting fraudulent sensor data to mislead control systems.	Grid instability or process failure.
Slow DoS	Network/ App	Stealthy low-volume traffic mimicking legitimate slow nodes.	Resource exhaustion with low visibility.
Adversarial Evasion	Learning	Manipulating inputs to fool ML models.	Compromised detection accuracy.

Sources: [3] [23] [36]

The fundamental vulnerability of IoT systems lies in the "security-by-design" gap. Devices are often deployed with default credentials, unpatched firmware, and a lack of standardized development guidelines. Furthermore, the reliance on lightweight protocols such as IEEE 802.15.4e or MQTT, while efficient for communication, can increase susceptibility to topological manipulation and resource exhaustion attacks [3] [23] [36].

TABLE II. Comparative Benchmarking of Surveyed MML Architectures for IoT Security.

Architecture	Fusion Strategy	Dataset	Acc%	F1	AUC
CNN-LSTM	Hybrid	Custom multi-modal	96,04 [37]	0.960	–
CNN-LSTM / CNN-GRU (IDS)	Hybrid	CICIoT2023 / ToN-IoT	>99.6 [36,37]	0.996	-
RansomFormer	Cross-attention	Custom ransomware	99.5 [43]	0.995	0.998
MT-CMVAD	Transformer fusion	UCF-Crime / CUHK	- [41]	-	0.989
TTGNet-AMD	Hybrid DL + GCN	Drebin / VirusTotal	~98.7 [30]	0.985	0.990
FW-CNN + Metaheuristic	Feature-level	ToN-IoT	98.4 [11]	0.982	-
Soft-Voting Ensemble	Decision-level	ToN-IoT / Bot-IoT	99.1 [49]	0.989	-
AI4FIDS (Federated)	Late fusion (FL)	Multi-domain IIoT	97.8 [40]	0.976	-
Autoencoder (Anomaly	Latent-space	N-BaIoT	97.3 [39]	0.971	0.981
GAN-augmented MML	Feature-level	CICIoT2023	99.3 [39]	0.991	-

Source: [11, 30, 36, 37, 39, 40, 41, 43, 49]

Accuracy and F1 values are reported on the primary dataset cited by each study (see Table II).

B. Overview of Data Modalities in IoT

In a multimodal context, a "modality" refers to a distinct stream of information that captures a specific aspect of a system's state or behavior. IoT ecosystems are rich in these modalities [12] [15] [16] [18] [32] [61]:

- Network Flow and Packets: This includes tabular features like flow duration, protocol type, and connection rates, as well as raw packet-level features such as payload entropy and application-layer semantics (e.g., MQTT message types).
- Device Telemetry: Internal performance metrics such as CPU load, RAM usage, memory mapping, and power consumption patterns.
- Sensor Data: The primary functional output of the device, which may include numerical readings (temperature, pressure) or high-dimensional data (images, audio).
- Static Code Features: Structural information extracted from software binaries, such as opcode sequences, Application Programming Interface (API) call sequences, and function call graphs (FCG).
- Behavioral Biometrics: Patterns derived from human interaction, including keystroke dynamics, gait, and touch gestures.

C. Foundations of Multimodal Machine Learning (MML)

Multimodal Machine Learning is the computational process of integrating information from multiple modalities to perform inference. The effectiveness of an MML system depends largely on its fusion strategy, the stage at which different data types are combined [12] [17] [18] [19] [61].

1) Early Fusion (Data-Level): This strategy involves concatenating raw features from different sources into a single vector before feeding it into the model. While this allows the model to capture deep inter-modal interactions from the beginning, it is highly sensitive to temporal misalignment and differences in data scales.

2) Late Fusion (Decision-Level): Each modality is processed by an independent model, and the individual outputs are merged at the final stage using techniques like majority voting, averaging, or weighted confidence scores. This is more robust to modality-specific noise but often misses the nuanced feature-level dependencies that define complex attacks.

3) Hybrid Fusion (Feature-Level): The most prevalent approach in modern research. It utilizes modality-specific encoders to extract high-level feature representations, which are then integrated in the intermediate layers of the network. This allows for a balance between modality-specific learning and global context integration.

III. Review Methodology

To ensure reproducibility and rigor, this review follows these systematic review steps.

A. Search Strategy

Literature was retrieved from major academic databases covering the period 2018 – 2026.

The digital libraries were especially, but not limited to: Google Scholar, IEEE Xplore, SpringerLink.

The following primary Boolean query was applied consistently across all selected databases:

("IoT" OR "IIoT") AND ("Multimodal Machine Learning" OR "Multi-modal Deep Learning" OR "Data Fusion" OR "Feature Fusion" OR "Cross-modal Attention") AND ("Cybersecurity" OR "Intrusion Detection" OR "Anomaly Detection" OR "Malware" OR "Authentication" OR "IDS" OR "Attack Detection")

B. Inclusion and Exclusion Criteria

Inclusion criteria: 1) Peer-reviewed articles or preprints from 2018-2026; 2) Explicit use of at least two distinct data modalities; 3) Application to IoT security (IDS, malware detection, or authentication); 4) Reporting of quantitative performance metrics.

Exclusion criteria: 1) Theoretical papers without experimental validation; 2) Single-modality studies; 3) non-English publications.

C. Sélection Process

The selection was based on the essential point of multimodality and intrusion detection systems, with papers addressing the most relevant databases and meeting the criterion of multimodality; the second selection criterion was recent references. This allowed us to keep about 60 reference articles.

IV. Taxonomy of MML Architectures for IoT Security

Recent progress in multimodal machine learning for IoT security has focused on creating systems that combine different types of data and make clear how they are all connected. In the contexts of IoT and IIoT, these modalities may include network traffic characteristics, device-level telemetry, system logs, sensor data, and, in some cases, behavioral signals. This section of the study groups the literature by the basic architecture of machine learning, with a focus on deep learning models, attention-based methods, Transformer-centric architectures, and hybrid ensemble frameworks. The goal is to look at how different types of architecture do multimodal fusion, deal with interactions between features from different modes, and solve basic IoT problems like data heterogeneity, temporal dynamics, and limited computing power [1] [7] [9] [34].

A. Deep Learning-Based Multimodal Models

Deep Learning (DL) is the most popular MML paradigm because it can learn hierarchical representations without manual feature engineering [18] [34] [35] [43] (see Table III).

TABLE III. MML ARCHITECTURE TAXONOMY

Architecture Family	Early Fusion	Hybrid Fusion	Late Fusion
CNN-based	Feature concatenation before CNN layers [35]	FW-CNN + metaheuristic hyperparameter tuning [11]	Ensemble: CNN + RF/SVM decision voting [34][48]
RNN/LSTM-based	Multimodal LSTM on concatenated sequences [37]	CNN-LSTM / CNN-GRU — spatial + temporal [6][35][37]	LSTM outputs + federated aggregation [8]
Transformer-based	Token-level cross-modal embedding [33]	RansomFormer cross-attention (bytes+API) [42]; MT-CMVAD [40]	Transformer ensemble with confidence weighting [34]
Hybrid Ensemble	Multi-feature static+dynamic fusion [25][29]	TTGNet-AMD: LSTM + Transformer-GCN [29]; AI4FIDS FL [39]	Soft-voting ensemble over independent models [11][48]
Autoencoder/ GAN	Joint latent-space reconstruction [10]	GAN augmentation + CNN classifier [38]	GAN synthetic data + downstream detector [10][38]

B. Convolutional and Recurrent Neural Network Integration

A common architecture for analyzing network and telemetry data is the combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). CNN layers are typically used to extract spatial or structural patterns from a window of data, while LSTM/GRU layers capture the temporal dependencies essential for identifying multi-stage attacks that unfold over time. The research studied highlights the efficacy of the hybrid CNN-LSTM model, which achieved an average accuracy of 96.04% across various behavioral modalities. In network intrusion detection, CNN-LSTM and CNN-GRU models have demonstrated accuracies exceeding 99.83%, surpassing the CNN model by 0.93% in accuracy, outperforming traditional deep learning techniques by effectively modeling both the spatial structure of packet flows and their sequential order [6] [35] [34] [37][62].

C. Autoencoders and Generative Models

Autoencoders (AE) are frequently employed in multimodal settings to learn a shared latent representation across different data sources. By training an AE to reconstruct "normal" multimodal inputs, any significant reconstruction error in a new input can be flagged as an anomaly. Generative Adversarial Networks (GANs) are increasingly used to enhance MML-based detection. GANs serve a dual purpose: first, they generate synthetic multimodal data to address the common problem of class imbalance, where malicious samples are rare compared to benign traffic, improving the robustness of the classifier. Second, they facilitate adversarial training, helping models recognize and resist "evasion" attacks where adversaries deliberately manipulate one modality to mask anomalies in another [10] [38] [41].

D. Attention and Transformer-Based MML

The most significant architectural shift in recent years is the adoption of the Transformer model and its attention mechanisms. Unlike traditional models that treat all inputs with equal weight, attention mechanisms allow a model to dynamically focus on the most relevant features or modalities for a specific task [10] [33] [40] [42].

1) Cross-modal attention mechanisms: The "cross-attention" mechanism is particularly vital for MML. It allows one modality (the "query") to search for relevant features in another modality (the "key" and "value"). This enables the model to align and bridge the semantic gap between disparate data types. For instance, in malware analysis, the RansomFormer model uses a cross-attention fusion layer to relate static Portable Executable (PE) byte data with dynamic API call sequences. The mathematical foundation of this fusion typically involves a query vector Q derived from one modality (e.g., bytes) and key K and value V vectors derived from another (e.g., APIs). The attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right) \cdot V$$

where, d_k is the dimensionality of the keys. This allows the model to prioritize API calls that are most relevant to specific byte-level signatures, achieving near-perfect accuracy (up to 99.5%) in ransomware detection [42].

Similar transformer-based models like MT-CMVAD have shown significant improvements in video anomaly detection, achieving AUC scores of up to 98.9% while reducing computational complexity [41].

2) Hybrid and ensemble models: Despite the dominance of deep learning, hybrid models that combine traditional machine learning (ML) with DL often provide better results in resource-constrained IoT settings. Traditional models like Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT) are prized for their simplicity, interpretability, and low computational requirements [2] [34] [48].

In a hybrid architecture, a deep learning model might be used for high-level feature extraction, with the final classification performed by a robust ensemble. For example, the FW-CNN model utilizes a one-dimensional CNN for feature learning, but tunes its hyperparameters using a hybrid metaheuristic optimizer to enhance its stability and detection rate. Another approach, the soft-voting ensemble classifier, fuses outputs from different specialized models using a confidence-weighted strategy, which has been shown to significantly reduce false positives in multimodal threat detection [11] [34] [48].

V. APPLICATIONS AND USE CASES

The practical impact of MML in IoT is best illustrated by its application to specific cybersecurity goals, where the fusion of diverse data sources directly addresses the weaknesses of unimodal systems [5] [6] [8] [39].

A. Multimodal Intrusion Detection Systems (M-IDS)

The dominant application of MML is the Intrusion Detection System (IDS). Modern M-IDS frameworks seek to create a "holistic" view of network and device behavior [5] [6] [8] [9] [35] [37].

1) Cyber-physical and industrial fusion: In smart grids and IIoT, M-IDS combines network traffic logs with physical sensor readings. An attacker might attempt a False Data Injection (FDI) to destabilize a power grid. While the individual sensor readings might appear within normal bounds, an MML model can detect the attack by identifying the subtle inconsistency between the sensors' reported states and the corresponding control-layer network commands [4] [5].

2) Federated and privacy-preserving detection: Systems like AI4FIDS leverage Federated Learning (FL) to combine data from multiple domains (e.g., different factories or hospitals), without compromising local data privacy. This allows the system to learn from a global pool of heterogeneous attack signatures while keeping sensitive local data behind its own firewall [8] [39].

3) Packet and metadata analysis: Traditional IDS often ignores the application layer. Newer MML models, such as the one proposed in NAIIDS4IoT, fuse low-level flow statistics with rich application-layer metadata, such as MQTT topic structures and payload entropy. This enables the detection of stealthy "Slow DoS" attacks that mimic legitimate network latency [5] [35].

B. Multimodal Malware Analysis

IoT malware has become increasingly evasive, using packing and encryption to hide its code from static scanners. MML addresses this by fusing static code features with dynamic behavior logs [10] [25] [28] [29] [43] [44] [52]. Table IV presents the multimodal malware analysis.

TABLE IV. MULTIMODAL MALWARE ANALYSIS

Feature Type	Specific Modalities	Core Advantage in MML
Static	Opcode sequences, Function Call Graphs (FCG), Byte-images.	Captures the comprehensive structural intent of the code.
Dynamic	System call traces, Network activity, Memory usage.	Resilient to code obfuscation and packing.
Fused	Cross-attention over structural and behavioral features.	Maximizes detection accuracy (99%+) and attribution to families.

Source: [10] [25] [28] [29] [43] [44] [52]

The TTGNet-AMD framework exemplifies this approach by using LSTM to process temporal opcode sequences while simultaneously using a Transformer-GCN hybrid to analyze the structural FCG. This collaborative interaction allows the model to identify malicious behavior even if the binary has been heavily obfuscated [29].

C. Secure Authentication and Access Control

Authentication in IoT is transitioning from static passwords to Continuous Authentication (CA) based on multimodal identity factors [26] [30] [31] [49] [50].

1) Behavioral and interaction fusion: MML models can verify a user's identity in real-time by fusing motion sensor data (accelerometer, gyroscope) with touch interaction patterns. This creates a "personal motion fingerprint" that is unique to the individual's biomechanics [31].

2) Biometric and liveness detection: Advanced Multi-Factor Authentication (MFA) systems integrate physiological traits (face, fingerprint, iris) with liveness indicators like Electrocardiogram (ECG) signals. Fusing these signals through a Siamese Neural Network architecture ensures that the system is not fooled by high-resolution photos or synthetic fingerprint replicas [30] [49] [50].

3) Continuous risk assessment: CA systems use MML to continuously monitor risk based on current and historical context. If a user's behavioral signature (e.g., typing speed or gait) deviates significantly from their profile, the system can automatically trigger a re-authentication request or lock the device [26] [31].

VI. DATASETS AND EVALUATION

The development of high-performing MML models is dependent on the availability of high-quality, diverse datasets that reflect the unique challenges of the IoT environment [13][24][27][46][47][51].

A. Common Multimodal IoT Security Datasets

Traditional datasets like NSL-KDD or Kyoto 2006+ are largely outdated for IoT research, as they do not capture the specific telemetry and protocols used in modern smart systems. Several benchmark datasets have emerged to fill this gap:

1) ToN-IoT: Sourced from the UNSW Canberra repository, this dataset integrates telemetry data from various sensors (weather stations, smart fridges) with network traffic logs and system event files from heterogeneous IoT devices. It captures a wide range of modern attacks, including ransomware, backdoors, and password cracking [13] [27].

2) Bot-IoT: A realistic dataset generated at the Cyber Range Lab of UNSW-Canberra, containing 72 million records that model IoT botnet activities such as DoS, DDoS, and information theft [51].

3) IoTID20: This dataset models a smart home network with devices like AI speakers, WiFi cameras, and smartphones, capturing 83 features across benign and malicious sessions [47].

4) CICIoT2023: An extensive and realistic dataset that executes 33 different attack types (divided into 7 classes) against a topology of 105 devices, including IoT devices as both attackers and victims [24].

5) N-BaIoT: Addressing the lack of public botnet data, this dataset provides real traffic from nine commercial IoT devices infected by Mirai and BASHLITE botnets [46] [51].

B. Evaluation Metrics and Benchmarking

Evaluating MML security systems requires a move beyond simple accuracy. In cybersecurity, the costs of a False Negative

(missing an attack) and a False Positive (blocking a legitimate user) are often asymmetrical.

1) Precision, Recall, and F1-score: These metrics are essential for evaluating performance on imbalanced datasets. Precision measures the reliability of an alarm, while Recall measures the system's ability to catch all malicious events. The F1-score provides the harmonic mean, which is critical for balancing these trade-offs.

2) Equal Error Rate (EER): Primarily used in authentication, EER identifies the point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal, serving as a primary indicator of a system's robustness.

3) Inference latency and resource overhead: For IoT devices, the time and energy required to make a prediction are as important as the prediction's accuracy. Metrics such as "packets per second", throughput, and model size in kilobytes (KB) are increasingly used to benchmark models for edge deployment [20] [21] [45].

VII. Edge-Cloud Deployment Trade-offs

Deploying Multimodal Machine Learning (MML) models on resource-constrained Internet of Things (IoT) devices necessitates a principled trade-off between detection accuracy and computational efficiency. Within this context, edge computing emerges as a foundational architectural pillar, enabling low-latency inference and reduced bandwidth consumption through localized data processing, while simultaneously reinforcing system resilience via adaptive, on-device responses to environmental stressors or adversarial conditions. Conversely, while cloud-centric platforms offer substantial advantages for large-scale model training, centralized data governance, and long-term policy management, they introduce non-trivial inference delays and inherit systemic vulnerabilities, including heightened privacy risks and exposure to single points of failure. To reconcile these competing imperatives, hybrid edge–cloud architectures distribute computational workloads based on task criticality: real-time anomaly flagging and time-sensitive inference are executed at the edge, whereas computationally intensive cross-modal analysis, periodic model retraining, and high-level coordination are selectively offloaded to the cloud. This collaborative paradigm establishes a complementary, rather than substitutive, relationship between edge and cloud tiers, thereby simultaneously satisfying the demands of real-time responsiveness, systemic scalability, and long-term operational resilience in resource-constrained IoT environments [20][21][45][53] (see Table V).

TABLE V. Quantitative Edge vs. Cloud Inference Trade-Offs

Deployment Mode	Inference Latency	Accuracy Δ -vs. Cloud	Representative Ref.
Full Cloud	50–200 ms	Baseline (0%)	[20][21]
Full Edge (quantised INT8)	5–15 ms	−2 to −4%	[45][58][60]
Selective Edge-Cloud Offload	8–25 ms (local); 80–150 ms (offloaded)	−0.5 to −1.5%	[20][21][53]
Federated (distributed edge)	10–30 ms per node	−1 to −3%	[8][39]

VIII. Open Challenges and Future Directions

The journey from academic research to practical, large-scale deployment of MML in IoT security is hindered by several significant obstacles that define the current research frontier [3] [20] [21] [22] [45] [53].

A. Data Synchronization and Spatiotemporal Alignment

A fundamental challenge in MML is the "asynchronous nature" of IoT data. Different sensors operate at different sampling rates; for instance, a network flow meter might provide updates every second, while a thermal sensor reports every minute, and a camera captures thirty frames per second.

Integrating these discordant streams requires robust alignment techniques. Current research into cross-modal attention mechanisms and transformer-based interpolation offers a path forward, allowing models to learn relationships across time steps even when data is missing or sporadically available. Furthermore, spatiotemporal misalignment where sensors have different spatial resolutions or perspectives remains an open problem, particularly in smart city environments [15] [16] [17] [32] [40].

B. Explainability and Trustworthiness (XAI)

As MML models grow in complexity, they increasingly suffer from the "black-box" issue, where their decision-making logic is opaque to human analysts. In security contexts, this lack of transparency is a critical liability, as security personnel must be able to validate and trust a system's output before taking remediation actions.

The integration of Explainable AI (XAI) tools like SHAP and LIME is becoming mandatory for advanced M-IDS. SHAP (Shapley Additive Explanations), based on game theory, can identify the specific contribution of each feature or modality to a prediction, providing a "visual explanation" that can help an analyst distinguish between a genuine cyberattack and a benign anomaly. Future research must focus on making these XAI methods computationally lightweight enough to run on IoT hardware [14] [41] [54] [55] [56] [57].

C. Resource Constraints and the Optimization Triad

The most persistent challenge is the "resource-performance trade-off". High-accuracy deep learning models, particularly large multimodal transformers, are computationally intensive and require significant memory and energy. IoT devices, however, often have power budgets below 10 watts and only kilobytes (KB) or megabytes of RAM [20] [21] [45] [53] [58] [59] [60].

The "Optimization Triad" addressing this includes:

1) Model compression: This includes pruning (removing redundant weights or neurons), quantization (reducing numerical precision from 32-bit floats to 8-bit integers), and knowledge distillation (training a small "student" model to mimic a large "teacher" model) [58] [59] [60].

2) Hardware acceleration: Development of specialized edge AI chips and hardware-aware pruning strategies that align

the model architecture with the specific processing units of the device [45] [53] [60].

3) Collaborative edge-cloud architectures: Implementing selective cloud offloading, where simple anomalies are processed locally at the edge for low latency, while complex, cross-modal analysis is offloaded to powerful cloud clusters [20] [21] [45] [53].

4) Toward antifragile and autonomous systems: Looking toward 2026 and beyond, the research community is moving toward "Antifragility", the design of systems that not only resist attacks (robustness) or recover from them (resilience) but actually grow stronger and more intelligent through exposure to stress and volatility. This involves integrating continuous streaming learning that can adapt to "concept drift" in real-time without human intervention. Furthermore, the rise of "Agentic AI" suggests a future where autonomous agents will coordinate defense responses across billions of devices, using multimodal insights to predict and mitigate threats before they even occur [3] [22] [61].

IX. Conclusion

The convergence of the Internet of Things and advanced machine learning has fundamentally redefined the parameters of cybersecurity. As this systematic review has demonstrated, the transition from unimodal to multimodal security architectures is not merely a technical upgrade but a necessary paradigm shift for defending complex, interconnected ecosystems. By integrating diverse and heterogeneous data streams—ranging from network flow statistics to behavioral biomechanics—Multimodal Machine Learning (MML) provides the holistic situational awareness required to detect and mitigate contemporary multi-vector threats [1] [3] [7] [22].

The classification of current literature reveals a clear trajectory toward deep learning models augmented by dynamic attention mechanisms. Architectures such as multimodal transformers and hybrid CNN-LSTM networks have established new benchmarks in accuracy for intrusion detection, malware analysis, and continuous authentication. These models effectively leverage the complementarity of different modalities, bridging the semantic gap between low-level telemetry and high-level behavioral patterns [10] [34] [35] [40] [42].

However, the field remains at a critical juncture. The challenges of data synchronization, model explainability, and resource constraints on the IoT edge must be addressed before MML can be deployed universally. The emerging optimization triad of model compression, hardware acceleration, and collaborative edge-cloud architectures provides a promising roadmap for future development. Ultimately, the goal is to move beyond passive defense toward autonomous, antifragile security systems that can self-heal and adapt to an ever-evolving threat landscape, ensuring the long-term safety, privacy, and resilience of our increasingly digitalized world [3] [20] [22] [45] [53] [61].

References

[1] Z. He, D. Davila, S. Bi, T. Wang, and T. Hou, "Machine Learning for Cybersecurity: A Survey of Applications, Adversarial Challenges, and Future Research Directions," *Electronics*, vol. 14, no. 23, p. 4563, Nov. 2025, doi: 10.3390/electronics14234563.

[2] A. Zahoor, W. Abbasi, M. Z. Babar, and A. Aljohani, "Robust IoT security using isolation forest and one class SVM algorithms," *Sci. Rep.*, vol. 15, no. 1, p. 36586, Oct. 2025, doi: 10.1038/s41598-025-20445-4.

[3] B. Alotaibi, "A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions," Dec. 19, 2025. doi: 10.20944/preprints202512.1717.v1./preprints202512.1717.v1 [Online]. Available: https://www.preprints.org/manuscript/202512.1717/v1.

[4] T. Bhuiyan, "AI in Smart Grid Cybersecurity: A Systematic Review of Machine Learning and Deep Learning Approaches against False Data Injection and Other Emerging Attacks," *JCSTS*, vol. 7, no. 8, pp. 1207–1295, Aug. 2025, doi: 10.32996/jcsts.2025.7.8.136.

[5] Y. S, "AI-Augmented Intrusion Detection Systems for Mitigating Advanced Persistent Threats in Cyber-Physical Manufacturing Networks," Oct. 06, 2025. doi: 10.20944/preprints202510.0470.v1./preprints202510.0470.v1 [Online].Available: https://www.preprints.org/manuscript/202510.0470/v1.

[6] S. I. Popoola, Y. Tsado, A. A. Ogunjinmi, E. Sanchez-Velazquez, Y. Peng, and D. B. Rawat, "Multi-Stage Deep Learning for Intrusion Detection in Industrial Internet of Things," *IEEE Access*, vol. 13, pp. 60532–60555, 2025, doi: 10.1109/ACCESS.2025.3557959.

[7] H. A. S. Ali and V. R. J, "Machine Learning for Internet of Things (IoT) Security: A Comprehensive Survey," *IJCNA*, vol. 11, no. 5, pp. 617–659, Oct. 2024, doi: 10.22247/ijcna/2024/40.

[8] M. Devine, S. P. Ardakani, M. Al-Khafajiy, and Y. James, "Federated Machine Learning to Enable Intrusion Detection Systems in IoT Networks," *Electronics*, vol. 14, no. 6, p. 1176, Mar. 2025, doi:10.3390/electronics14061176.

[9] H. Liao *et al.*, "A survey of deep learning technologies for intrusion detection in Internet of Things," *IEEe Access*, vol. 12, pp. 4745–4761, 2024, doi: 10.1109/ACCESS.2023.3349287.

[10] T. Nazmin, "A Dynamic Hierarchical Attention Framework for Multimodal Malware Detection," Aug. 2025, [Online]. Available: https://scholarworks.uark.edu/etd/5933 [Online]. Available: https://www.osti.gov/servlets/purl/2584221.

[11] H. A. Alsalamah and W. N. Ismail, "Evolutionary Computation for Feature Optimization and Image-Based Dimensionality Reduction in IoT Intrusion Detection," *Mathematics*, vol. 13, no. 23, p. 3869, Dec. 2025, doi: 10.3390/math13233869.

[12] V. Singh, "Multimodal deep learning: Integrating text, vision, and sensor data: Developing models that can process and understand multiple data modalities simultaneously," International Journal of Research in Information Technology and Computing. https://romanpub. com/ijaetv4-1-2022. php, vol. 4, no. 1, Jun. 2022, Accessed: Jan. 21, 2026. [Online]. Available: https://romanpub.com/resources/Vol.%204%20No.%201%20(June%2C%202022)%20-%2035.pdf

[13] Z. Cao, Z. Zhao, W. Shang, S. Ai, and S. Shen, "Using the ToN-IoT dataset to develop a new intrusion detection system for industrial IoT devices," *Multimedia Tools and Applications*, vol. 84, no. 16, pp. 16425–16453, 2025.

[14] İ. Kök, F. Y. Okay, Ö. Muyanlı, and S. Özdemir, "Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14764–14779, Aug. 2023, doi: 10.1109/JIOT.2023.3287678.

[15] T. Sadiq and C. W. Omlin, "Sensing in Smart Cities: A Multimodal Machine Learning Perspective," *Smart Cities*, vol. 9, no. 1, p. 3, Dec. 2025, doi: 10.3390/smartcities9010003.

[16] M. J. C. S. Reis, "Internet of Things and Artificial Intelligence for Secure and Sustainable Green Mobility: A Multimodal Data Fusion Approach to Enhance Efficiency and Security," *MTI*, vol. 9, no. 5, p. 39, Apr. 2025, doi: 10.3390/mti9050039.

[17] Q. Ge, C. Wei, and D. Wu, "Abnormal Network User Detection Method based on the Hybrid Neural Network from the Perspective of Multimodal Data Fusion," *J CIRCUIT SYST COMP*, vol. 34, no. 08, p. 2550195, May 2025, doi: 10.1142/S0218126625501956.

[18] A. Saghir, A. Akbar, A. Hasan, and A. Zafar, "Deep learning for multi-modal data fusion in IoT applications," *Mehran University Research*

Journal of Engineering & Technology, vol. 44, no. 1, pp. 75–81, 2025, doi: 10.22581/muet1982.3171.

[19] A. Kate, "Multimodal Machine Learning for Financial Fraud Detection Across Banking and E-Commerce Ecosystems," 2025.

[20] X. Wang and W. Jia, "Optimizing edge ai: a comprehensive survey on data, model, and system strategies.(2025)," *arXiv preprint arXiv:2501.03265*.

[21] S. A. Cajas Ordóñez, J. Samanta, A. L. Suárez-Cetrulo, and R. S. Carbajo, "Intelligent Edge Computing and Machine Learning: A Survey of Optimization and Applications," *Future Internet*, vol. 17, no. 9, p. 417, Sep. 2025, doi: 10.3390/fi17090417.

[22] B. Alotaibi, "A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions," Dec. 19, 2025. doi: 10.20944/preprints202512.1717.v1.

[23] B. Monteiro and J. Granjal, "Trust-Aware Distributed and Hybrid Intrusion Detection for Rank Attacks in RPL IoT Environments," *IoT*, vol. 7, no. 1, p. 4, 2025, doi: 10.3390/iot7010004.

[24] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," *Sensors*, vol. 23, no. 13, p. 5941, Jun. 2023, doi: 10.3390/s23135941.

[25] S. Wang and Y. Wang, "Malware Classification Technology Combining Multimodal Fusion with Deep Learning Algorithms," *Journal of Cyber Security and Mobility*, vol. 14, no. 3, pp. 597–622, 2025, doi: 10.13052/jcsm2245-1439.1434.

[26] B. Alotaibi and M. Alotaibi, "Hybrid Deep Learning Framework for Continuous User Authentication Based on Smartphone Sensors," *Sensors*, vol. 25, no. 9, p. 2817, Apr. 2025, doi: 10.3390/s25092817.

[27] Moustafa, Nour (TON-IoT Network Intrusion Dataset,). [Online]. Available: https://research.unsw.edu.au/projects/toniot-datasets

[28] S. Maganur, Y. Jiang, J. Huang, and F. Zhong, "Feature-Centric Approaches to Android Malware Analysis: A Survey," *Computers*, vol. 14, no. 11, p. 482, Nov. 2025, doi: 10.3390/computers14110482.

[29] J. Feng, M. Wang, J. Song, C. Yang, L. Zhang, and Y. Shen, "TTGNet - AMD: Android malware detection based on multi-modal feature fusion," *PeerJ Computer Science*, vol. 11, p. e3412, Dec. 2025, doi: 10.7717/peerj-cs.3412.

[30] P. P. Jena, K. N. Kattigenahally, S. Nikitha, S. Sarda, and H. Y, "Multimodal Biometric Authentication: Deep Learning Approach," in *2021 International Conference on Circuits, Controls and Communications (CCUBE)*, Bangalore, India: IEEE, Dec. 2021, pp. 1–5. doi: 10.1109/CCUBE53681.2021.9702724.

[31] S. K. Natarajan, A. Abdullah, S. Kaur, and P. Natarajan, "Advancing Multi-Modal Behavioral Biometric Authentication: A Deep Learning Approach with Synthetic Data Generation," *IEEE Access*, 2025.

[32] C. Wang, Y. Xu, and D. Zhang, "Multi-Modal Perception and Fusion for Maritime Autonomy: A Survey," 2025, doi: 0.20944/preprints202510.0977.v1.

[33] R. K. Solanki, A. M. Dhore, and A. R. Gadekar, "AUGMENTING AR/VR EXPERIENCES USING MULTIMODAL TRANSFORMER MODELS," *Int. J. of Electronics Engineering and Applications*, vol. 13, no. 2, 2025, [Online]. Available: https://ijeea.in/wp-content/uploads/15_21.pdf

[34] A. Odeh and A. Abu Taleb, "Ensemble-based deep learning models for enhancing IoT intrusion detection," *Applied Sciences*, vol. 13, no. 21, p. 11985, 2023, doi: 10.3390/app132111985.

[35] K. O. Adefemi, M. B. Mutanga, and O. A. Alimi, "A Hybrid CNN–GRU Deep Learning Model for IoT Network Intrusion Detection," *Journal of Sensor and Actuator Networks*, vol. 14, no. 5, p. 96, 2025, doi: 10.3390/jsan14050096.

[36] S. K. R. Mallidi and R. R. Ramisetty, "Optimizing Intrusion Detection for IoT : A Systematic Review of Machine Learn8ng and Deep Learning Approaches With Feature Selection and Data Balancing," *WIREs Data Min & Knowl*, vol. 15, no. 2, p. e70008, Jun. 2025, doi: 10.1002/widm.70008.

[37] L. Zhen, N. H. Kamarudin, V. J. Kok, and F. Qamar, "Anomaly detection model in network security situational awareness based on machine learning: Limitation, techniques, future trends," *IEEE Access*, 2025. **DOI:** 10.1109/ACCESS.2025.3589620

[38] N. Peppes, E. Daskalakis, T. Alexakis, and E. Adamopoulou, "A Multimodal Framework for Advanced Cybersecurity Threat Detection Using GAN-Driven Data Synthesis," *Applied Sciences*, vol. 15, no. 15, p. 8730, 2025. doi.org/10.3390/app15158730

[39] P. Radoglou-Grammatikis *et al.*, "AI4FIDS: Multimodal Federated Intrusion Detection," *IEEE Trans. Emerg. Topics Comput.*, pp. 1–15, 2025, doi: 10.1109/TETC.2025.3562346.

[40] H. Ding, S. Lou, H. Ye, and Y. Chen, "MT-CMVAD: A Multi-Modal Transformer Framework for Cross-Modal Video Anomaly Detection," *Applied Sciences*, vol. 15, no. 12, p. 6773, 2025, [Online]. Available: https://doi.org/10.3390/app15126773

[41] J. Dong *et al.*, "A Novel Multimodal Data Fusion Framework: Enhancing Prediction and Understanding of Inter-State Cyberattacks," *BDCC*, vol. 9, no. 3, p. 63, Mar. 2025, doi: 10.3390/bdcc9030063.

[42] S. Alzahrani, Y. Xiao, S. Asiri, N. Alasmari, and T. Li, "ansomFormer: A Cross-Modal Transformer Architecture for Ransomware Detection via the Fusion of Byte and API Features," *Electronics*, vol. 14, no. 7, p. 1245, Mar. 2025, doi: 10.3390/electronics14071245.

[43] B. Alotaibi, "Multimodal Deep Learning Fusion for Accurate and Explainable Malware Family Classification," *Applied Sciences*, vol. 15, no. 21, p. 11635, Oct. 2025, doi: 10.3390/app152111635.

[44] M. U. Tanveer, K. Munir, H. J. Alyamani, S. R. Hassan, M. Sheraz, and T. C. Chuah, "Graph-augmented multi-modal learning framework for robust android malware detection," *Scientific Reports*, vol. 15, no. 1, p. 38341, 2025.

[45] R. Kumar and A. Sharma, "Edge AI: a review of machine learning models for resource-constrained devices," *Artificial Intelligence and Machine Learning Review*, vol. 5, no. 3, pp. 1–11, 2024.

[46] F. Taher, M. Abdel-Salam, M. Elhoseny, and I. M. El-Hasnony, "Reliable machine learning model for IIoT botnet detection," *Ieee access*, vol. 11, pp. 49319–49336, 2023. doi: 10.1109/ACCESS.2023.3253432

[47] O. B. J. Rabie, S. Selvarajan, T. Hasanin, A. M. Alshareef, C. K. Yogesh, and M. Uddin, "A novel IoT intrusion detection framework using Decisive Red Fox optimization and descriptive back propagated radial basis function models," *Scientific Reports*, vol. 14, no. 1, p. 386, 2024.doi.org/10.1038/s41598-024-51154-z

[48] P. Zhao, "Distributed SVM-based Multimodal Intrusion Detection Architecture With Incremental Learning And Modality Scalability," *IJCAI*, vol. 49, no. 7, Nov. 2025, doi: 10.31449/inf.v49i7.9350.

[49] A. Ganmati, K. Afdel, and L. Koutti, "Deep Learning-Based Multi-Factor Authentication: A Survey of Biometric and Smart Card Integration Approaches," 2025, *arXiv*. doi: 10.48550/ARXIV.2510.05163.

[50] V. Velmurugan and S. Priyadharshni, "AI-Driven Secure Authentication: A Deep Learning Approach for Multi-Modal Biometric Systems," in *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, IEEE, 2025, pp. 49–56.

[51] Y. Meidan, et al. "detection_of_IoT_botnet_attacks_N_BaIoT," UCI Machine Learning Repository, 2018. [Online]. Available: https://doi.org/10.24432/C5RC8J.

[52] Z. Qian *et al.*, "Evaluating Diverse Feature Extraction Techniques of Multifaceted IoT Malware Analysis: A Survey," *arXiv e-prints*, p. arXiv: 2509.03442, 2025.

[53] S. O. Semerikov, T. A. Vakaliuk, O. B. Kanevska, O. A. Ostroushko, and A. O. Kolhatin, "Edge intelligence unleashed: a survey on deploying large language models in resource-constrained environments," *Journal of Edge Computing*, vol. 4, no. 2, pp. 179–233, 2025.

[54] J. Moss, J. Gordon, W. Duclos, Y. Liu, Q. Wang, and J. Wang, "Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation," *Electronics*, vol. 14, no. 23, p. 4622, 2025.

[55] K. S. Alketbi and A. Mehmood, "A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection," *IEEE Access*, 2025.

[56] G. Rjoub *et al.*, "A survey on explainable artificial intelligence for cybersecurity," *IEEE Transactions on Network and Service Management*, vol. 20, no. 4, pp. 5115–5140, 2023.

[57] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, "Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM," *Plos one*, vol. 20, no. 5, p. e0318542, 2025, doi: 10.1371/journal.pone.0318542.

[58] R. AGRAWAL and H. KUMAR, "Energy-Efficient AI Model Design for Edge Devices Using Neural Network Pruning and Optimization Techniques," 2023.

[59] J. Greg, 'Quantization and Pruning Strategies for Energy-Efficient TinyML Models', vol. Volume 6 Issue 10, 2025.

[60] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2022, doi: 10.1109/JPROC.2022.3226481.

[61] P. Berrang, "Custom Software Development Blog," 7T.ai, 2025. [Online]. Available: https://7t.ai/digital-transformation-insights-blog/.

[62] M. Latif *et al.*, 'AI-based intelligent sensing detection of cybersecurity threats using multimodal sensor data in smart devices', *Scientific Reports*, vol. 16, no. 1, p. 11091, Feb. 2026, doi: 10.1038/s41598-026-40614-3.

Optimizing the Multi-Omics Data Types and Variant Combinations for Accurate Breast Cancer Molecular Subtypes Classification

Sajid Shah[1], Azurah A Samah[2]*, Syed Hamid Hussain Madni[3],
Sarina Sulaiman[4], Zuraini Ali Shah[5], Wong Yee Leng[6], Aryati Bakri[7]
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia [1, 2, 4, 5, 6, 7]
School of Electronics and Computer Science, University of Southampton Malaysia, 79100 Iskandar Puteri, Johor, Malaysia [3]

***Abstract*—Breast cancer is a highly heterogeneous disease with Luminal-A, Luminal-B, HER2-Enriched, Basal-Like, and Normal-Like molecular subtypes. Accurate classification of breast cancer molecular subtypes is essential for effective diagnosis, treatment, and planning. In recent years, multi-omics data has been widely used to improve classification performance. However, most of the existing studies focus on various combinations of multi-omics data types and variants without considering their biological relevance and computational effectiveness. This research study aims to systematically analyze, validate, and optimize the combinations of multi-omics data types and variants for accurate breast cancer molecular subtypes classification. The main goal is to identify the most suitable biologically meaningful combinations for improving classification performance. This research study provides the biological rationale for integrating the multi-omics data types and variants, and analyzes the various combinations used by existing studies for breast cancer subtype classification and the reasons behind their selection. Based on this analysis, possible and best-proposed combinations of multi-omics data types and variants are presented for the accurate classification of breast cancer molecular subtypes, based on both biological and computational perspectives. In addition, this research study identifies and recommends reliable public databases that provide multi-omics datasets with verified PAM50 labels for accurate subtype classification. The findings can help researchers design more accurate and reliable classification models by using the best proposed combination of multi-omics data types and variants, and select appropriate datasets with validated subtype labels.**

***Keywords*—Breast cancer; molecular subtypes; multi-omics; data integration; data types and variants; datasets**

I. Introduction

Breast cancer is considered one of the most serious health issues worldwide. According to GLOBOCAN [1], Kim et al. [2], the breast cancer cases recorded 2.6 million new cases and 685,000 deaths in 2024, and breast cancer became the most commonly diagnosed cancer. To treat the patients, doctors need to know the exact type of breast cancer and help them receive the fastest and best treatment, knowing the condition of breast cancer [3].

The genetic alterations and DNA damage can be influenced by estrogen exposure in the cell, that led to breast cancer. DNA defects or cancer-causing genes like BRCA1 and BRCA2 can occasionally be inherited [4]. Consequently, the chance of developing breast cancer is increased if breast cancer has already appeared in the family. In a healthy person, the immune system attacks the cells with aberrant DNA or abnormal development. When breast cancer patients experience this failure, tumors develop and spread [5].

Researchers have found that breast cancer is not one single disease. It has different molecular subtypes that behave in different ways, as shown in Fig. 1. Based on the heterogeneity, breast cancer is classified into two stratifications, which are histological and molecular [6]. Histological classification is mainly based on morphological study, whereas molecular classification is based on the molecular pattern of breast cancer. Knowing the specific molecular subtype is very important because each subtype responds and plays differently in the process of treatment [6].

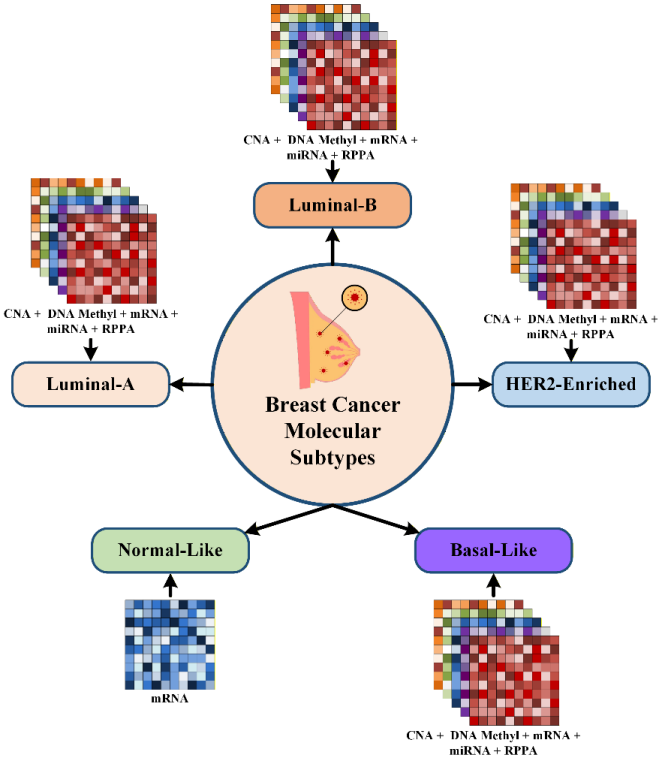


Fig. 1. Breast cancer molecular subtypes

*Corresponding author

Despite all this progress, existing research studies for breast cancer molecular subtypes classification lack accurate combinations of multi-omics data types and variants. The various combinations of data types and variants are used for breast cancer molecular subtypes classification without thinking deeply about the biology behind them, such as how genes, proteins, and chemical marks work together in each subtype. Both biological reasons (for example, which layers best capture hormone pathways or cell growth signals) and computer science reasons (for example, which mixes give stable models and clear results) play an important role in the classification of breast cancer molecular subtypes.

According to Shah et al. [5], multi-omics can improve accuracy, but there are still problems that exist, including HDLSS, limited explainability of AI models, class imbalance, multi-omics integration, and choosing the correct multi-omics data types and variant combinations. Many existing studies select various combinations of multi-omics data types and variants, but they do not explain why those combinations are best for the breast cancer molecular subtypes classification. Also, the majority of the breast cancer classification studies are struggling to identify verified and authentic databases for breast cancer datasets, while most of the studies also find it hard to locate the datasets that are completed with accurate PAM50 subtypes labeling.

Wrong or incomplete classification of breast cancer molecular subtypes due to the inappropriate combination of data types and variants, selection of wrong databases and datasets can lead to bad treatment choices and poorer patient results. Therefore, there is a pressing need to identify the optimal multi-omics data type and variant combinations along with the verified multi-omics databases and datasets for the accurate and reliable classification of breast cancer molecular subtypes. It will help create more reliable tools for doctors and move the field toward truly personalized medicine for breast cancer.

The main contributions of this research study to the field of breast cancer research and multi-omics data analysis are mentioned below:

- It provides a comprehensive investigation of various combinations of multi-omics data types and variants for breast cancer molecular subtype classification based on biological and computational evidence.
- It clearly explains the biological rationale for integrating multi-omics data, showing the biological importance of combining these multi-omics data types and variants necessary for accurate classification.
- It analyzes the existing studies on multi-omics integration for breast cancer subtype classification and highlights the various multi-omics data types and variant combinations.
- It proposes the optimal possible and best combination of multi-omics data types and variants based on both strong biological evidence and computational effectiveness for breast cancer molecular subtypes classification.
- It identifies and recommends reliable public databases and datasets that provide complete multi-omics datasets with accurate PAM50 subtype labels for future research.

The rest of the study is organized as follows: Section II emphasizes the breast cancer molecular subtypes, while Section III focuses on the omics, including mono-omics and multi-omics, along with the details of the multi-omics data types and variants. Section IV is about the biological rationale for combining multi-omics data types in breast cancer molecular subtypes. Furthermore, Section V describes multi-omics integration for breast cancer molecular subtype classification. Section VI is about the database and dataset selection. Section VII is future directions and suggestions and lastly, Section VIII is the conclusion of the study.

II. Breast Cancer Molecular Subtypes

Breast cancer is categorized into the five molecular subtypes, including the Luminal-A, Luminal-B, HER2-Enriched, Basal-Like and Normal-Like. Table I shows the description of the breast cancer molecular subtypes based on gene expression features, growth rate and treatment prognosis. Different breast cancer molecular subtypes are identified by unique gene expression patterns that affect on tumors behave and tumors respond to treatment [7]. Each molecular subtype of breast cancer has been elaborated in detail in the subsections below:

A. Luminal-A

Luminal-A is one of the most common molecular subtypes of breast cancer and about 50% to 60% of all breast cancer cases. It is characterized by the higher expression of the ER(+) and PR(+) and it has a HER2(-) status [8]. This tumor has a lower level of proliferation-related genes (measured by Ki-67). Breast cancer patients with Luminal-A have often better prognoses as compared to the other molecular subtypes, including lower recurrence risk and higher survival rates [9].

B. Luminal-B

Luminal-B occur for 15% to 20% of breast cancer molecular subtypes and are often more aggressive. The hormone-positive molecular of Luminal-B is characterized by ER(+), PR(+) and HER2(+) status [8]. As compared to Luminal-A, Luminal-B tumors typically show high proliferative [10]. Also, it shows more aggressive clinical and biological behaviour.

C. HER2-Enriched

HER2 occur 15% to 20% of the breast cancer molecular subtypes. It is characterized by high expression of the HER2(+) and protein along with the ER(-) and PR(-) status [8]. This molecular subtype confers more aggressive clinical and biological behavior.

D. Basal-Like

Basal-Like is also known as Triple-Negative Breast Cancer (TNBC) and is characterized by the absence of ER(-), PR(-) and HER2(-) [8]. This tumor is more aggressive than the Luminal-A and Luminal-B and grows faster. The genes of these tumors are typically found in the basal/myoepithelial cells of the abnormal breast [11]. It affects younger women, particularly those who have BRCA1 gene mutations.

TABLE I. DESCRIPTION OF BREAST CANCER MOLECULAR SUBTYPES

Subtypes	Gene Expression Features	Growth Rate	Prognosis	Percentage
Luminal-A	ER(+), PR(+), HER2(-)	Slow	Normal	About 50% to 60%
Luminal-B	ER(+), PR(+), HER2(+)	Fast	Bad	About 15% to 20%
HER2-Enriched	ER(-), PR(-), HER2(+)	Faster	Worse	About 15% to 20%
Basal-Like	ER(-), PR(-), HER2(-)	Fastest	Worst	About 8% to 37%
Normal-Like	ER(- /+), PR(-/+), HER2 (-/+)	Slow to Moderate	Good	About 5% to 10%

E. Normal-Like

Normal-Like accounts for about 5% to 10% of all breast cancer molecular subtypes. Normal-Like is characterized by ER(-), PR(-) and HER2(-) status [8]. It is a rare molecular subtype that is identified with the help of gene expression profiling that demonstrates similar characteristics to normal breast tissues. Normal-Like typically expresses the genes that are associated with the adipose tissues and other non-epithelial cells that are found in the breast [8].

III. OMICS DATA

Omics data refers to the large-scale biological data that contain various levels of molecular information of a living organism. The term "omics" contains biological information such as genomics (study of DNA and genetic variations), epigenomics (study of heritable changes in gene expression that do not involve changes in the DNA sequence), transcriptomics (study of RNA and gene expression) and proteomics (study of proteins) [12]. Each omics data type provides a unique structure and view of the biological organism [13].

A. Mono-Omics

Mono-omics data is the utilization of a single type of omics data type to study biological systems, such as genomics or epigenomics or transcriptomics. Mono-omics analysis is simpler, easier to process and requires less computational power compared to multi-omics analysis [14]. Mono-omics data types are widely used in previous research and have offered insights into disease mechanisms. For instance, Cristovao, et al. [15], Zeng, et al. [16], Li and Nabavi [17] use mon-omics data, particularly on genomics or transcriptomics data types, to analyze and classify breast cancer molecular subtypes. However, mono-omics data focus on capturing only one layer of biological information, which limits the potential of mono-omics data to fully represent the complexity of heterogeneous diseases such as breast cancer molecular subtypes classification [18].

B. Multi-Omics

Multi-omics data refers to the integration of various data types of omics to provide a detailed and comprehensive structure and view of biological systems [19]. Instead of depending on a single layer of information, such as mono-omics data type, multi-omics combines different molecular data sources at different molecular layers of an organism to better understand the complex and heterogeneous diseases, including the breast cancer molecular subtypes classification [20]. Multi-omics data helps in capturing the important relationships and interactions among different biological processes of an organism that lead to more reliable and accurate classification of breast cancer molecular subtypes [21].

C. Multi-Omics Data Types

There are four main data types of multi-omics, including genomics, epigenomics, transcriptomics and proteomics. Each multi-omics data type is described in subsections.

1) Genomics: It is the study of the complete set of DNA within a cell, including all genes and their functions. It concentrates on the understanding of genetic changes in an organism, such as insertions, deletions and mutations [22]. In breast cancer molecular subtypes, genomics assists in identifying the acquired or inherited genetic alterations that trigger cancer development or influence tumor behavior in the breast [23]. The genetic changes in breast cancer deactivate or activate the oncogenes and tumor suppressor genes that lead to uncontrolled breast cancer cell growth [22]. Overall, genomics provides the foundational layer of biological information for understanding the origin of breast cancer at the DNA level.

2) Epigenomics: It is the study of changes in gene activity that do not involve alterations in the DNA sequence itself. Instead, it concentrates on the regulatory mechanisms such as DNA-Methylation, histone modification and chromatin structure changes. The epigenetic modifications can turn genes "on" or "off" any gene without changing the underlying genetic code [24]. In breast cancer, abnormal epigenetic changes lead to the silencing of important tumor suppressor genes or the activation of harmful genes, contributing to breast cancer progression [25]. Epigenomics is important because it contains the information about the environmental and biological factors that can influence gene behavior without changing DNA.

3) Transcriptomics: It is the study of RNA molecules that are produced from DNA through the process of transcription. It mainly concentrates on the gene expression levels, demonstrating which genes are transcribed actively into RNA under specific conditions [26]. In breast cancer, transcriptomics assists in identifying the differences in gene expression patterns across different cancer molecular subtypes, such as Luminal or Basal-like types [27]. These expression patterns provide important insights into tumor activity, cellular processes, and disease progression.

4) Proteomics: Proteomics is the study of proteins, which are the final functional products of genes. It concentrates on the protein abundance, modifications, interactions and structure within biological systems [28]. Proteins carry out most cellular

functions; proteomics provides direct insight into the way cells behave in normal and disease conditions [29]. In breast cancer, abnormal protein interactions or expression indicate tumor metastasis, growth and treatment resistance [27]. Proteomics is important because it reflects the actual functional state of the cell, making it highly valuable for understanding disease mechanisms and improving subtype classification.

Overall, the integration of multi-omics data types is very important for breast cancer molecular subtypes classification. Each omics layer provides complementary and unique information, and combining them helps capture the full complexity of breast cancer biology. By integrating multi-omics data types in an appropriate biological and computational way, researchers can improve classification accuracy, identify more reliable biomarkers, and develop better personalized treatment strategies for different breast cancer subtypes.

D. Multi-Omics Data Variants

In breast cancer research, multi-omics data are further categorized into different variants that represent specific types of molecular measurements within each multi-omics data type. These variants provide more detailed and fine-grained biological information about tumor development, progression, and subtype differences [30]. Instead of treating each omics layer as a single dataset, researchers analyze different variants such as genetic alterations, gene expression levels, epigenetic modifications, and protein activity. Each multi-omics data variant captures a unique aspect of cancer biology, and when combined, they help to build a more complete and accurate model for breast cancer molecular subtype classification [31].

1) Copy Number Alteration (CNA): Copy Number Alteration refers to changes in the number of copies of a particular DNA segment in the genome. These alterations can include gains (amplifications) or losses (deletions) of chromosomal regions. CNA is a key variant of genomics because it directly reflects structural changes in DNA that may lead to cancer development by increasing oncogene activity or deleting tumor suppressor genes [32].

2) Copy Number Variation (CNV): Copy Number Variation is a form of genomic structural variation where sections of the genome vary in copy number between individuals. CNVs can influence gene dosage and are strongly associated with cancer susceptibility and tumor progression. In breast cancer, CNVs help explain variability in gene expression and subtype-specific genomic instability [32].

3) Single Nucleotide Polymorphism (SNP): Single Nucleotide Polymorphisms are variations at a single nucleotide position in DNA. SNPs are the most common type of genetic variation and can affect how genes function or how proteins are produced. In breast cancer genomics, SNPs are used to identify genetic risk factors and subtype-specific mutations [33].

4) DNA Methylation: DNA methylation is an epigenomic modification where methyl groups are added to DNA, usually affecting gene expression without changing the DNA sequence. It plays a major role in gene silencing and activation [34]. In breast cancer, abnormal DNA methylation patterns are linked to tumor suppression loss and subtype-specific regulatory changes [35].

5) Messenger RNA Expression (mRNA): mRNA represents the transcriptomic level of gene activity, measuring how actively genes are being transcribed into RNA. It reflects real-time cellular function and is widely used for defining breast cancer intrinsic subtypes, such as Luminal and Basal-Like. mRNA profiles provide strong signals for tumor classification and biological interpretation [36].

6) MicroRNA Expression miRNA: miRNAs are small non-coding RNA molecules that regulate gene expression by binding to target mRNA and preventing protein production. miRNA is considered a regulatory transcriptomic variant that influences cancer progression by controlling multiple genes simultaneously [37]. In breast cancer, miRNA expression patterns are associated with subtype differences and tumor aggressiveness [38].

7) Reverse Phase Protein Array (RPPA): Reverse Phase Protein Array is a proteomics-based technique used to measure protein expression and activation levels, including post-translational modifications such as phosphorylation. RPPA provides functional information about signaling pathways in breast cancer. It reflects the final stage of gene expression and is essential for understanding how molecular changes translate into cellular behavior and treatment response [39].

IV. Biological Rationale for Combining Multi-Omics Data Types in Breast Cancer Molecular Subtypes

Breast cancer is a highly heterogeneous disease with various molecular subtypes. Each subtype shows different genetic mutations, gene expression patterns, epigenetic regulation, and protein activity [23]. Mono-omics data types are not sufficient to fully explain tumor behavior because cancer development is driven by multiple biological mechanisms acting together across different omics levels. For instance, Basal-like tumors show very high TP53 mutation rates, while Luminal tumors show frequent PIK3CA and GATA3 mutations, and HER2-enriched tumors are strongly associated with ERBB2 amplification [39]. These differences clearly indicate that breast cancer cannot be fully understood using only one type of omics data.

A. Relationship of Multi-Omics Data Types and Variants

Genomics, epigenomics, transcriptomics, and proteomics multi-omics data types are closely associated and work together in order to describe the detailed and comprehensive view of a full biological system of a cell. Multi-omics data types are not independent; instead, they form an organized, sequenced, and layered biological flow where each multi-omic layer influences the next [40]. Genomics represents the foundation because DNA contains all the genetic information required for the development and function of a cell. However, the activity of a gene is not determined by DNA alone. Epigenomics acts as a regulatory layer on top of genomics by controlling whether specific genes are turned on or off through mechanisms. The epigenetic changes directly influence transcriptomics because only genes that are activated at the epigenomic level are transcribed into RNA. Transcriptomics then reflects the real-time gene expression profile of a cell, showing which genes are

actively producing RNA molecules under certain conditions. The RNA is further translated into proteins, making proteomics the final functional layer of the biological system. Proteins carry out most cellular activities, including signaling, metabolism, and structural functions, and therefore represent the actual biological outcome of genomic instructions [12]. The relationship among the multi-omics data types is highly interconnected, meaning that changes in one layer directly lead to effects in others, as shown in Fig. 2. Because of the strong interdependencies, integrating multi-omics data is very important for understanding heterogeneous and complex diseases like breast cancer. The interconnected biological flow enables researchers to capture not only individual molecular changes but also the dynamic relationships between different molecular layers, leading to more accurate and biologically meaningful cancer subtype classification [39].

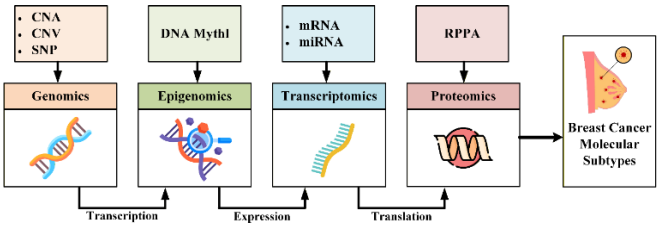


Fig. 2. Relationship among the multi-omics data types and variants.

B. Biological Dependency of Multi-Omics

One of the most important biological reasons for integrating multi-omics data is the strong dependency between multi-omics data types. DNA mutations can influence epigenetic regulation, which then affects gene expression and ultimately protein production. For example, TP53 mutations in Basal-like tumors lead to uncontrolled cell proliferation, while PIK3CA mutations in Luminal tumors activate downstream signaling pathways affecting both RNA expression and protein signaling [39]. Similarly, epigenetic silencing of tumor suppressor genes can reduce RNA and protein expression even when DNA remains unchanged. This cascading biological relationship shows that each omics layer only provides partial information, but together they describe the complete disease mechanism [39].

C. Importance of Multi-Omics Integration for Breast Cancer Molecular Subtypes Classification

Because breast cancer subtypes are defined by combined genetic, epigenetic, transcriptional, and proteomic features, integrating all these data types is necessary for accurate classification. TCGA multi-platform studies have shown that basal-like tumors are consistently distinct across all omics layers, while luminal and HER2 subtypes show overlapping patterns across different data types [41]. This proves that single-omics approaches fail to capture subtype complexity. Multi-omics integration allows researchers to identify more robust biomarkers, detect hidden subtype relationships, and improve classification accuracy by combining complementary biological signals [42].

In summary, breast cancer is driven by interconnected molecular mechanisms across multiple biological layers. Genomics explains structural changes, epigenomics explains regulatory control, transcriptomics reflects gene activity, and proteomics captures functional outcomes. However, none of these layers alone can fully represent tumor behavior. Therefore, integrating the genomics, epigenomics, transcriptomics, and proteomics multi-omics data is biologically necessary to capture the complete molecular architecture of breast cancer and improve the accuracy of molecular subtype classification.

V. Multi-Omics Integration for Breast Cancer Molecular Subtype Classification

The successful integration of multi-omics variants provides deep insights into the molecular alterations across various regulatory levels of breast cancer. The correct selection of multi-omics data types and variants is very important for breast cancer molecular subtype classification and gives us a full picture of the molecular landscape of breast cancer. This helps in identifying subtypes and finding possible biomarkers for personalised treatment plans [31]. There are various combinations of multi-omics data types and variants that are used by existing research studies for performing the breast cancer molecular subtypes classification, as shown in Fig. 3.

A. Combinations of Multi-Omics Data Types for Breast Cancer Molecular Subtypes Classification

Multi-omics data types integration has become very important in classifying the breast cancer molecular subtype. By combining genomics, epigenomics, transcriptomics, and proteomics multi-omics data types, researchers can capture the heterogeneous and complex structure and mechanism of breast cancer. As compared to mono-omics integration, multi-omics integration improves the ability to identify subtypes of specific biomarkers, discover the interactions and relationships among various multi-omics data types and variants, and enhance the overall classification and prediction performance.

Existing studies have used various combinations of multi-omics data types, as shown in Table II, to improve the breast cancer molecular subtypes classification. The existing studies use combinations of two or three data types to classify the breast cancer molecular subtypes.

TABLE II. Multi-Omics Data Types Used by the Existing Studies

No of Multi-Omics Data Types	Genomics	Epigenomics	Transcriptomics	Proteomics	References
2	✓		✓		[15-17]
2		✓	✓		[43-47]
3	✓	✓	✓		[48-55]
3	✓	✓		✓	[7]
3	✓		✓	✓	[56]
4	✓	✓	✓	✓	**Proposed**

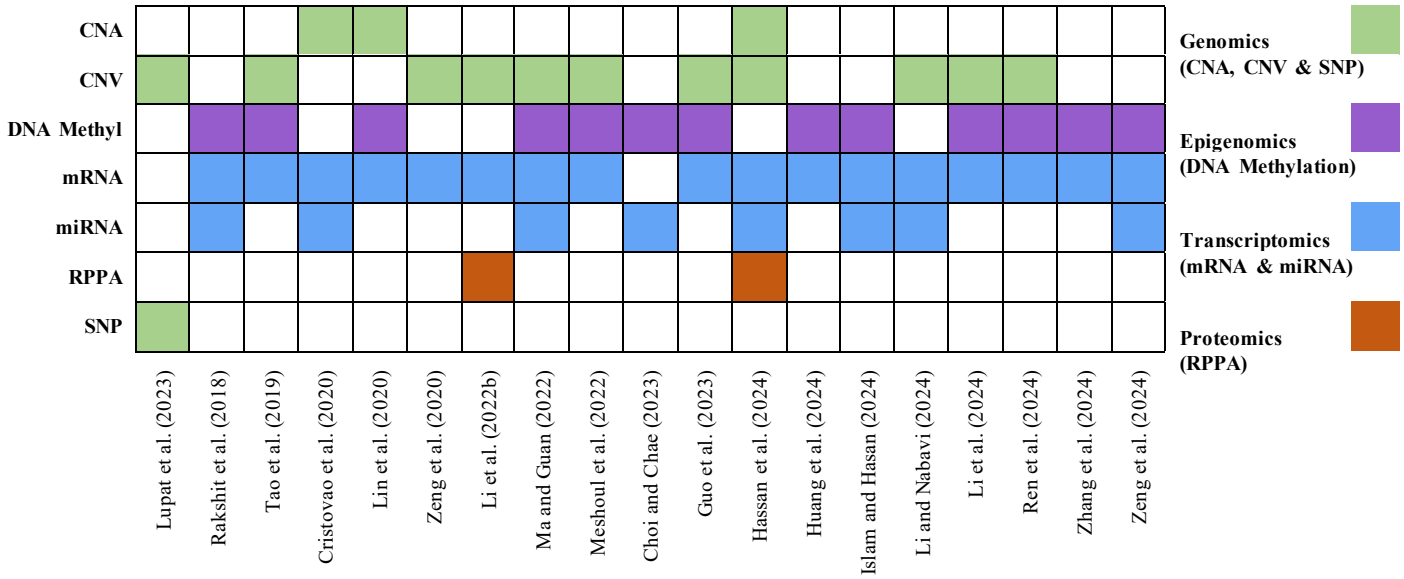


Fig. 3. Multi-omics data types and variants used by the existing studies.

Each combination of multi-omics data types provides unique insights; however, the existing studies that integrated three data types generally show significant performance. Based on the detailed analysis, it is suggested that the integration of four main multi-omics data types tends to provide a more detailed molecular structure, which can lead to more reliable and accurate classification of breast cancer molecular subtypes.

B. *Combinations of Multi-Omics Variants for Breast Cancer Molecular Subtypes Classification*

With the rapid growth of multi-omics data, Copy Number Variation (CNV), Single Nucleotide Polymorphism (SNP), DNA methylation, mRNA expression, miRNA expression, and protein abundance (RPPA), multi-omics variants have become available. The proper and correct integration of multi-omics variants allows for a more robust and reliable classification of breast cancer molecular subtypes by capturing the important information in each omics data type [42]. Researchers use various combinations of the multi-omics variants to improve the classification performance in breast cancer molecular subtypes classification.

Table III presents the various combinations of multi-omics variants that are utilized by the existing studies for the successful classification of breast cancer subtypes. Existing studies use combinations of two, three, four, or five multi-omics data variants for the classification of molecular breast cancer subtypes. Overall, the integration of multi-omics variants empowers researchers to explore and capture the complex molecular interactions to understand the mechanism and underlying structure of breast cancer molecular subtypes. Each multi-omics variant combination provides unique information about the breast cancer subtypes on the molecular level. Based on the detailed analysis, it is suggested that integrating the five multi-omics variants can enhance breast cancer molecular subtypes classification accuracy and performance.

TABLE III. MULTI-OMICS DATA VARIANTS USED BY THE EXISTING STUDIES

No of Multi-Omics Data Variants	CNA	CNV	SNP	DNA-M	mRNA	miRNA	RPPA	References
2		✓	✓					[57]
2		✓			✓			[16, 46]
2				✓	✓			[51]
3		✓		✓	✓			[49, 52-54]
3		✓		✓		✓		[48, 55]
3	✓				✓	✓		[15, 17]
3	✓				✓		✓	[7]
3				✓	✓	✓		[43]
4		✓		✓	✓	✓		[50]
5	✓	✓		✓	✓	✓		[56]
5		✓		✓	✓	✓	✓	**Proposed**

Notably, models incorporating the accurate multi-omics data types and variants tend to achieve superior classification performance and emphasize the value of multi-omics data fusion in breast cancer molecular subtypes analysis.

C. Proposed Combinations of Multi-Omics Variants for Breast Cancer Molecular Subtypes Classification Based on Biological Rationale

The reliable and appropriate integration of multi-omics data types and variants plays an important role in enhancing the accuracy and robustness of breast cancer molecular subtype classification. Based on biological evidence and insights from existing computational classification models, various combinations of multi-omics data variants are utilized for capturing the heterogeneous and complex view of breast cancer, as mentioned in Section V-A and Section V-B. These combinations are selected to represent different layers of biological information, including genomic alterations, epigenetic regulation, gene expression, and protein activity, which collectively define the breast cancer behavior and subtype characteristics.

From a biological perspective, the combination of genomics, epigenomics, transcriptomics, and proteomics multi-omics data types, along with their desired variants, allows for a comprehensive representation of breast cancer structure, behavior, and progression across multiple molecular levels.

Based on the biological rationale and computational models for breast cancer molecular subtypes classification, the combination of CNV, DNA Methylation, mRNA, and RPPA multi-omics data variants, which integrates genomic structure, epigenetic control, gene expression, and protein activity, or the combination of CNV, DNA methylation, miRNA, and RPPA multi-omics data variants, which replaces mRNA with miRNA, resulting in a combination of, focusing more on regulatory transcriptomic mechanisms. These possible combinations have demonstrated effectiveness in capturing subtype-specific patterns and improving classification performance compared to mono-omics analysis. However, these combinations are still incomplete and can be further explored by adding more multi-omics variants.

To address these limitations, this research study proposes an enhanced and more comprehensive combination that integrates CNV, DNA Methylation, mRNA, miRNA, and RPPA multi-omics data variants. The proposed combination is based on the biological justification because it captures all major layers of molecular regulation, starting from DNA-level structural variations (CNV), followed by epigenetic regulation (DNA-Methylation), gene expression (mRNA), gene regulation (miRNA) and finally protein-level functional outcomes (RPPA), as shown in Fig. 4. By including both mRNA and miRNA, the model benefits from a more complete representation of transcriptomic activity, including both expression and regulatory mechanisms. The inclusion of RPPA further strengthens the model by incorporating protein-level validation of upstream molecular changes.

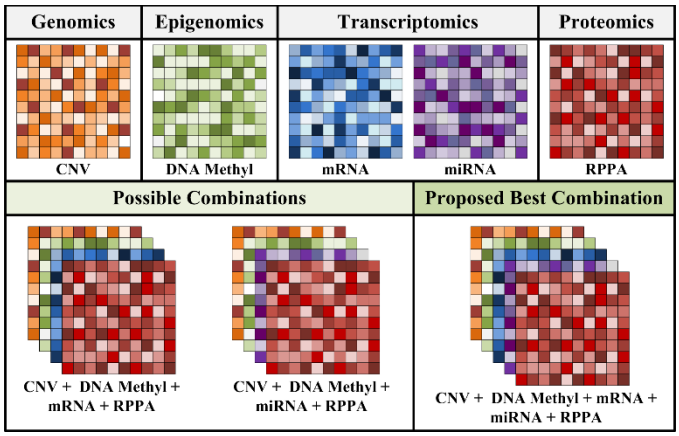


Fig. 4. Possible and proposed best combinations of multi-omics variants for breast cancer molecular subtypes classification.

From a computational perspective, integrating these complementary data variants can improve feature diversity, reduce information loss, and enhance the ability of machine learning and deep learning models to learn complex patterns associated with different breast cancer subtypes. The proposed comprehensive combination not only improves classification accuracy but also increases model robustness and biological interpretability. Therefore, the proposed combination of multi-omics data types and variants is considered the most suitable and effective strategy for accurate breast cancer molecular subtype classification.

VI. Database and Dataset Selection

There are various databases used for breast cancer research, as mentioned in Table IV, each with its own characteristics, quality, multi-omics types, and variants. Compared to other commonly used databases such as CPTAC, NCBI, GEO, and METABRIC, the TCGA and UCSC Xena databases are among the most widely used and reliable sources for breast cancer multi-omics datasets. TCGA (The Cancer Genome Atlas) is the original and one of the most comprehensive cancer research projects, led by the National Cancer Institute (NCI), which profiled more than 20,000 primary tumors across 33 different cancer types, including breast cancer. It provides a complete range of multi-omics data types and their variants, making it highly suitable for molecular subtype classification studies. The official repository for TCGA data is the Genomic Data Commons (GDC), where both raw and processed datasets are available. On the other hand, UCSC Xena is another powerful data visualization and exploration platform that serves as a user-friendly portal for accessing breast cancer datasets. UCSC Xena re-processes and organizes multi-omics data into easy-to-use formats, allowing researchers to efficiently explore, integrate, and download datasets across all major omics types and variants. Due to their comprehensive coverage of multi-omics data types and variants, as shown in Table IV, standardized processing, and availability of clinically annotated information such as PAM50 subtype labels, TCGA, and UCSC Xena are considered the most suitable and commonly used databases for breast cancer multi-omics research and classification tasks.

Table IV presents a widely used databases and repositories for breast cancer molecular subtype classification datasets. It shows that multi-omics data, particularly genomics, epigenomics, transcriptomics, and proteomics, can be effectively obtained from platforms such as Linkedomics-TCGA-BRCA, UCSC-XENA-BRCA, and GDC-TCGA-BRCA. While other repositories like CPTAC, NCBI, GEO, and METABRIC also provide valuable breast cancer datasets, they have certain limitations, such as missing omics layers (e.g., proteomics or epigenomics), restricted cohorts, or static data availability. Table V describes the multi-omics GDC-TCGA-BRCA dataset used for breast cancer molecular subtype classification. This dataset is downloaded from the GDC-TCGA repository and carefully preprocessed to ensure accurate subtype labeling using PAM50 standards. Since labeling inconsistencies are a common challenge in existing studies, all samples were manually validated and assigned standardized labels to improve reliability. The dataset includes multiple multi-omics data variants along with their respective feature dimensions and sample sizes.

TABLE IV. COMPARISON OF USED MULTI-OMICS DATASETS FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Datasets and Repositories	Genomics	Epigenomics	Transcriptomics	Proteomics	Limits
Linkedomics (Linkedomics-TCGA-BRCA) https://linkedomics.org/data_download/TCGA-BRCA/	Y	Y	Y	Y	N/A
University of California, Santa Cruz (UCSC-XENA-BRCA) https://xena.ucsc.edu	Y	Y	Y	Y	N/A
Genomic Data Commons – (GDC-TCGA-BRCA) https://portal.gdc.cancer.gov/projects/TCGA-BRCA	Y	Y	Y	Y	N/A
Clinical Proteomic Tumor Analysis Consortium (CPTAC) https://cptac-data-portal.georgetown.edu/cptacPublic/	N	N	N	Y	Only the proteomics cohort from TCGA
National Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/datasets/	Y	Y	Y	N	Proteomics missing
Gene Expression Omnibus (GEO) https://www.ncbi.nlm.nih.gov/geo/	Y	Y	Y	N	A cohort from NCBI missing proteomics Data
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Y	N	Y	N	Static data & genomics and proteomics are missing

TABLE V. DESCRIPTION OF MULTI-OMICS GDC-TCGA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	19569	759	754
Epigenomics	DNA-Methylation	19050	759	754
Transcriptomics	mRNA	18207	759	754
	miRNA	369	759	754

TABLE VI. DESCRIPTION OF MULTI-OMICS LINKEDOMICS-TCGA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	24776	1080	1039
Epigenomics	DNA-Methylation	20106	783	772
Transcriptomics	mRNA	20155	1093	1051
	miRNA	823	755	740
Proteomics	RPPA	175	887	858

TABLE VII. DESCRIPTION OF MULTI-OMICS UCSC-XENA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	60624	1082	1021
Epigenomics	DNA-Methylation	486428	893	774
Transcriptomics	mRNA	60661	1226	1052
	miRNA	1882	1202	1036
Proteomics	RPPA	488	919	853

Table VI outlines the multi-omics LINKEDOMICS-TCGA-BRCA dataset, which was obtained from the Linkedomics platform. Similar to the GDC dataset, this dataset was also rigorously preprocessed, and PAM50 labels were assigned and validated to ensure consistency and accuracy in breast cancer molecular subtype classification. It contains a more comprehensive set of omics data, including genomics, epigenomics, transcriptomics, and proteomics (RPPA), making it particularly suitable for integrative multi-omics analysis. Finally, Table VII presents the multi-omics UCSC-XENA-BRCA dataset, which was collected from the UCSC Xena repository. This dataset was also preprocessed following the same standardized pipeline, where PAM50 labels were carefully validated and assigned to address common labeling challenges in breast cancer datasets. It includes extensive high-dimensional data across genomics, epigenomics, transcriptomics, and proteomics, further supporting robust multi-omics-based classification. Collectively, these datasets provide a strong foundation for developing and evaluating accurate breast cancer molecular subtype classification models.

In addition, the PAM50-validated labels generated and refined in this research study for all the datasets in Table V, Table VI, and Table VII are publicly available to support reproducibility and future research. The breast cancer molecular subtypes PAM50-labels can be downloaded from the GitHub repository, enabling other researchers to utilize standardized and reliable data for breast cancer molecular subtypes classification.

VII. Suggestions and Future Directions

This research study shows that the best combination of multi-omics data types and variants for breast cancer molecular subtypes classification is CNV + DNA methylation + mRNA + miRNA + RPPA because it covers all important biological layers from DNA changes to protein activity. Future research should evaluate the proposed best combination with more advanced machine learning and deep learning models to determine whether classification accuracy and robustness can be improved further. Researchers can also explore additional possible and optimized combinations of multi-omics data types and variants to strengthen predictive performance and improve generalization capability.

In addition, the optimized multi-omics combinations identified and validated in this research study should be further tested through practical implementation and real-world validation using a breast cancer molecular subtype classification experiment. Practical simulation using actual multi-omics breast cancer datasets and clinical classification environments is important to confirm the reliability, scalability, and effectiveness of the identified combinations in real diagnostic scenarios.

It is also important to evaluate the proposed models on other cancer-related datasets and in clinical settings to assess their adaptability and clinical usefulness. Future work should further focus on improving model interpretability so that researchers and healthcare professionals can better understand the classification decisions. Moreover, challenges such as missing data handling, high dimensionality, data imbalance, and computational complexity should also be addressed in future studies. Furthermore, future studies should also investigate the computational cost and dimensionality challenges associated with integrating multiple high-dimensional multi-omics variants simultaneously. The integration of CNV, DNA methylation, mRNA, miRNA, and RPPA data can significantly increase data dimensionality, storage requirements, computational complexity, and model training time, particularly when implemented using standard computing resources. Therefore, future work should focus on developing computationally efficient frameworks, dimensionality reduction strategies, optimized feature selection methods, and scalable architectures to improve practical implementation feasibility and resource efficiency for large-scale multi-omics classification systems.

By following these directions, researchers can develop more reliable, interpretable, and accurate multi-omics classification systems for breast cancer molecular subtypes, which may contribute to better personalized treatment strategies for breast cancer patients as well as patients affected by other cancer types.

VIII. Conclusion

This research study investigated various combinations of multi-omics data types and variants for accurate breast cancer molecular subtype classification. It explained omics data, including mono-omics, multi-omics data types and variants, and also examined the biological reasons for combining the multi-omics data types, particularly genomics, epigenomics, transcriptomics, and proteomics, along with the justification of computational modelling studies that have performed breast cancer molecular subtypes classification using various combinations of multi-omics data. According to our observation, the majority of existing research does not choose combinations carefully based on both biology and computer science. To fill this gap, we proposed the best combination of CNV + DNA methylation + mRNA + miRNA + RPPA multi-omics variants to classify the breast cancer molecular subtypes because it captures all important layers of cancer biology and works well with classification models. Furthermore, this research study also recommended TCGA and UCSC Xena as the most suitable databases because they contain all the required multi-omics data types and variants with correct PAM50 subtype labels. This research study gives clear guidelines for researchers to build better, more accurate, and biologically meaningful classification models. Overall, the findings will help improve early diagnosis, personalized treatment, and patient outcomes in breast cancer.

Acknowledgment

This study was supported by the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UTM/03/1) from the Malaysian Ministry of Higher Education.

Data Availability

GitHub Repository Link: https://github.com/sajidshah232/Breast-Cancer-Molecular-Subtypes-PAM50-Labels.git

References

[1] GLOBOCAN, "Global Cancer Observatory " 2025.

[2] J. Kim et al., "Global patterns and trends in breast cancer incidence and mortality across 185 countries," Nature Medicine, pp. 1-9, 2025.

[3] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," Biological research, vol. 50, no. 1, p. 33, 2017.

[4] A. R. Venkitaraman, "How do mutations affecting the breast cancer genes BRCA1 and BRCA2 cause cancer susceptibility?," DNA repair, vol. 81, p. 102668, 2019.

[5] S. Shah, A. A. Samah, S. H. H. Madni, S. Z. M. Hashim, and M. Faheem, "Recent advancements in artificial intelligence-driven breast cancer molecular subtypes classification using multi-omics: A comprehensive review," Engineering Applications of Artificial Intelligence, vol. 170, p. 114237, 2026.

[6] R. G. do Nascimento and K. M. Otoni, "Histological and molecular classification of breast cancer: what do we know?," Mastology, vol. 30, pp. 1-8, 2020.

[7] X. Li et al., "MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis," Frontiers in Genetics, vol. 13, p. 806842, 2022.

[8] O. Yersal and S. Barutca, "Biological subtypes of breast cancer: Prognostic and therapeutic implications," World journal of clinical oncology, vol. 5, no. 3, p. 412, 2014.

[9] J. J. Gao and S. M. Swain, "Luminal a breast cancer and molecular assays: a review," The oncologist, vol. 23, no. 5, pp. 556-565, 2018.

[10] F. Ades et al., "Luminal B breast cancer: molecular characterization, clinical management, and future perspectives," Journal of clinical oncology, vol. 32, no. 25, pp. 2794-2803, 2014.

[11] J. Y. So, J. Ohm, S. Lipkowitz, and L. Yang, "Triple negative breast cancer (TNBC): Non-genetic tumor heterogeneity and immune microenvironment: Emerging treatment options," Pharmacology & therapeutics, vol. 237, p. 108253, 2022.

[12] U. Srivastava, S. Kanchan, M. Kesheri, M. K. Gupta, and S. Singh, "Types of omics data: genomics, metagenomics, epigenomics, transcriptomics, proteomics, metabolomics, and phenomics," in Integrative omics: Elsevier, 2024, pp. 13-34.

[13] S. M. Meystre, R. Gouripeddi, and A. V. Alekseyenko, "Molecular, Genetic, and Other Omics Data," in Clinical Research Informatics: Springer, 2023, pp. 309-328.

[14] J. Lee, D. Y. Hyeon, and D. Hwang, "Single-cell multiomics: technologies and data analysis methods," Experimental & Molecular Medicine, vol. 52, no. 9, pp. 1428-1442, 2020.

[15] F. Cristovao et al., "Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 1, pp. 121-134, 2020.

[16] J. Zeng, H. Cai, and T. Akutsu, "Breast cancer subtype by imbalanced omics data through a deep learning fusion model," in Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics, 2020, pp. 78-83.

[17] B. Li and S. Nabavi, "A multimodal graph neural network framework for cancer molecular subtype classification," BMC bioinformatics, vol. 25, no. 1, p. 27, 2024.

[18] A. Peng, X. Mao, J. Zhong, S. Fan, and Y. Hu, "Single-cell multi-omics and its prospective application in cancer biology," Proteomics, vol. 20, no. 13, p. 1900271, 2020.

[19] S. Graw et al., "Multi-omics data integration considerations and study design for biological systems and disease," Molecular omics, vol. 17, no. 2, pp. 170-185, 2021.

[20] S. Shah et al., "Multi-Omics Integration Methods for AI-Based Breast Cancer Molecular Subtypes Classification," International Journal of Advanced Computer Science & Applications, vol. 17, no. 1, 2026.

[21] G. T. Jung, K.-P. Kim, and K. Kim, "How to interpret and integrate multi-omics data at systems level," Animal cells and systems, vol. 24, no. 1, pp. 1-7, 2020.

[22] A. M. Lesk, Introduction to genomics. Oxford University Press, 2017.

[23] U. Testa, G. Castelli, and E. Pelosi, "Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments," Medical Sciences, vol. 8, no. 1, p. 18, 2020.

[24] K. Gagnidze and D. W. Pfaff, "Epigenetic mechanisms: DNA methylation and histone protein modification," in Neuroscience in the 21st century: from basic to clinical: Springer, 2022, pp. 2677-2716.

[25] R. Mathur et al., "Epigenetic factors in breast cancer therapy," Frontiers in Genetics, vol. 13, p. 886487, 2022.

[26] J. Rajawat, "Transcriptomics," in Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data: Springer, 2019, pp. 39-56.

[27] Z. Zhu, L. Jiang, and X. Ding, "Advancing breast cancer heterogeneity analysis: insights from genomics, transcriptomics and proteomics at bulk and single-cell levels," Cancers, vol. 15, no. 16, p. 4164, 2023.

[28] C. Monti, M. Zilocchi, I. Colugnat, and T. Alberio, "Proteomics turns functional," Journal of proteomics, vol. 198, pp. 36-44, 2019.

[29] T. D. Veenstra, "Omics in systems biology: current progress and future outlook," Proteomics, vol. 21, no. 3-4, p. 2000235, 2021.

[30] D. Lee, Y. Park, and S. Kim, "Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches," Briefings in bioinformatics, vol. 22, no. 3, p. bbaa188, 2021.

[31] M. M. Ortiz and E. R. Andrechek, "Molecular characterization and landscape of breast cancer models from a multi-omics perspective," Journal of mammary gland biology and neoplasia, vol. 28, no. 1, p. 12, 2023.

[32] G. Hovhannisyan, T. Harutyunyan, R. Aroutiounian, and T. Liehr, "DNA copy number variations as markers of mutagenic impact," International journal of molecular sciences, vol. 20, no. 19, p. 4723, 2019.

[33] K. S. Allemailem et al., "Single nucleotide polymorphisms (SNPs) in prostate cancer: its implications in diagnostics and therapeutics," American journal of translational research, vol. 13, no. 4, p. 3868, 2021.

[34] G. A. Dhar, S. Saha, P. Mitra, and R. Nag Chaudhuri, "DNA methylation and regulation of gene expression: Guardian of our health," The Nucleus, vol. 64, no. 3, pp. 259-270, 2021.

[35] K. Holm et al., "Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns," Breast cancer research, vol. 12, no. 3, p. R36, 2010.

[36] A. Szymiczek, A. Lone, and M. R. Akbari, "Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review," Clinical genetics, vol. 99, no. 5, pp. 613-637, 2021.

[37] K. Pierouli et al., "Long non-coding RNAs and microRNAs as regulators of stress in cancer," Molecular medicine reports, vol. 26, no. 6, p. 361, 2022.

[38] K. Kalecky, R. Modisette, S. Pena, Y.-R. Cho, and J. Taube, "Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including miR-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes," BMC cancer, vol. 20, no. 1, p. 141, 2020.

[39] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, and J. McMichael, "Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012; 490: 61–70," ed, 2021.

[40] R. Duan et al., "Evaluation and comparison of multi-omics data integration methods for cancer subtyping," PLoS computational biology, vol. 17, no. 8, p. e1009224, 2021.

[41] P. Goel and N. Shaikh, "Uncovering Clinically Relevant Breast Cancer Subtypes Biomarkers Using Integrative Bioinformatics and Machine Learning Approaches," Biomarkers, no. just-accepted, pp. 1-12, 2026.

[42] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," Computational and structural biotechnology journal, vol. 19, pp. 949-960, 2021.

[43] S. Rakshit, I. Saha, S. S. Chakraborty, and D. Plewczyski, "Deep learning for integrated analysis of breast cancer subtype specific multi-omics data," in TENCON 2018-2018 IEEE region 10 conference, 2018, pp. 1917-1922: IEEE.

[44] Y. Huang, P. Zeng, and C. Zhong, "Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning," BMC bioinformatics, vol. 25, no. 1, p. 132, 2024.

[45] S. Islam and M. N. Hasan, "Personalized graph feature-based multi-omics data integration for cancer subtype identification," arXiv preprint arXiv:2408.08832, 2024.

[46] Q. Zhang et al., "AET-net: A framework for subtype classification based on the multi-omics data of breast cancer," Molecular & Cellular Biomechanics, vol. 21, no. 4, pp. 785-785, 2024.

[47] P. Zeng, C. Huang, and Y. Huang, "DiffRS-net: A Novel Framework for Classifying Breast Cancer Subtypes on Multi-Omics Data," Applied Sciences, vol. 14, no. 7, p. 2728, 2024.

[48] M. Tao et al., "Classifying breast cancer subtypes using multiple kernel learning based on omics data," Genes, vol. 10, no. 3, p. 200, 2019.

[49] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," Genes, vol. 11, no. 8, p. 888, 2020.

[50] Y. Ma and J. Guan, "MOCSC: a multi-omics data-based framework for cancer subtype classification," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2022, pp. 2853-2859: IEEE.

[51] J. M. Choi and H. Chae, "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks," BMC bioinformatics, vol. 24, no. 1, p. 169, 2023.

[52] H. Guo, X. Lv, Y. Li, and M. Li, "Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification," Briefings in Functional Genomics, vol. 22, no. 5, pp. 463-474, 2023.

[53] N. Li, F. Yang, Q. Zhang, Q. Li, X. Zhang, and J. Teng, "CautionGCN: Cancer Subtype Classification by Developing Causal Multi-Head Autoencoder and Graph Convolutional Network," in 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI), 2024, pp. 626-634: IEEE.

[54] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable multi-class classification based on integrative feature selection for breast cancer subtyping," Mathematics, vol. 10, no. 22, p. 4271, 2022.

[55] Y. Ren et al., "Classifying breast cancer using multi-view graph neural network based on multi-omics data," Frontiers in Genetics, vol. 15, p. 1363896, 2024.

[56] A. M. Hassan, S. M. Naeem, M. A. Eldosoky, and M. S. Mabrouk, "Multi-omics-based Machine Learning for the Subtype Classification of Breast Cancer," Arabian Journal for Science and Engineering, pp. 1-14, 2024.

[57] R. Lupat, R. Perera, S. Loi, and J. Li, "Moanna: multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes," Ieee Access, vol. 11, pp. 10912-10924, 2023.

An AI-Driven Framework for Software Effort Estimation Based on Developer Performance Metrics

Shaheer Ahmed[1], Nosheen Qamar[2], Faria Nazir[3], Nosheen Sabahat[4], Atif Ikram[5]*, Najla Abdulaziz Almousa[6], Hebah Abdullah Abubakr[7], Mohammed Abual-Rub[8], Abdulrahman Alojail[9], Marwan Abu-Zanona[10]
Department of Software Engineering, University of Management and Technology, Lahore, Pakistan [1, 2, 3]
Department of Computer Science, Forman Christian College University, Lahore, Pakistan[4]
Department of Computer Science, University of Lahore, Lahore, Pakistan[5]
College of Information Technology, Amman Arab University, Amman, 11118, Jordan[5]
Department of Management Information Systems-College of Business Administration,
King Faisal University, Al-Ahsa 31982, Saudi Arabia[6, 7, 8, 9, 10]

***Abstract*—Effort estimations, including budgets, hiring people, and project timelines, in the Agile methodology, are determined by tools like COCOMO and Function-Point analysis. This study presents a framework driven by artificial intelligence (AI) that uses almost real-time signals from Git platforms that track issues, and tools to determine code quality, convert them into vectors, and trains four different regressors on them: ordinary least-squares regression, a random-forest ensemble, gradient-boosted trees, and a long short-term memory network. Hold-out evaluation together with five-fold cross-validation supplies mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination, complemented by feature-importance charts from the tree-based learners. A CI/CD-integrated retraining schedule keeps the estimator aligned with evolving team dynamics. Analyzing multi-developer projects over successive sprints reveals where predictions remain accurate and where unpredictable behavior emerges, pointing to chances for improved data gathering, enhanced governance, and more intentional feature development.**

Keywords*—Software effort estimation; machine learning; agile development; artificial intelligence; software analytics; developer performance

I. Introduction

Forecasting how much effort a software task will consume is central to budgeting, headcount decisions, and release planning. For a long time, the industry leaned on parametric formulas, such as COCOMO and its variants, function-point analysis, and related calibration tools that convert size proxies and cost drivers into person-month estimates [1]. Those methods still dominate plan-driven shops because they are transparent and fit neatly into historical cost-accounting workflows [2] [3]. Surveys of estimation practice show that these approaches remain common in plan-driven contexts because they are transparent and align with historical cost accounting [2].

Teams usually tune these formulas with local multipliers, but the tuning data go stale fast when the underlying technology changes, team members rotate out, or coding standards get overhauled. The staleness that results is most visible when an organization pivots to continuous delivery and suddenly needs short-range forecasts rather than the multi-year roadmaps it had been using [4] [5].

Agile practices compress planning horizons and introduce constant scope flux, which makes it unrealistic to expect that one locked-in parameter set will reliably capture a team's capacity from sprint to sprint. Review after review of agile estimation surfaces the same issues: estimates that miss the mark, over-confident forecasts, and a stubborn preference for gut feeling, even when historical logs go untouched [6]. On the machine-learning side, a different picture emerges. Support-vector machines, multilayer perceptrons, generalized linear models, and tree ensembles have all been shown to push error down on repository-scale datasets whose features capture size and complexity. AI-focused surveys in software-engineering cost prediction single this out as one of the most economically consequential forecasting challenges in the discipline [4] [5]. ML provides alternative algorithms, including SVM, multilayer perceptrons, generalized linear models, and tree ensembles, which can cut error on repository-style datasets once the feature capture size and complexity are well enough [7] [8] [9].

Once the developer metrics change, even a bit, they affect the speed of the project [10]. The practical outcome is that forecasts start to diverge from realized velocity whenever individual throughput, review workload, or collaboration habits shift [5] [11].

Nobody has fully tapped the continuous stream of signals from version-control systems, issue trackers, quality gates, and team-chat tools as first-class inputs inside a repeatable prediction loop. Practices like planning poker, bucket sizing, and affinity grouping remain useful for alignment, but they scale poorly and cannot absorb fresh telemetry on their own. When a prediction is missed, more research on costing takes priority [10] [12]. To avoid that, manual approaches, like planning poker and bucket sizing, take place. But since they are difficult to scale and do not absorb fresh inputs, automatically [13].

With that backdrop, our study focuses on an internal environment where Git, Jira, SonarQube, pull-request platforms, and messaging tools are already mandatory. We set out to derive sprint-level features, benchmark several regressors side by side, and show how scheduled retraining inside a CI/CD pipeline can keep predictions in step with a changing team [14].

Four research questions steer the investigation: RQ1 : How can real-time developer metrics sharpen agile effort estimates?

*Corresponding author

RQ2: Which engineered metrics carry the most predictive weight? RQ3: Among the four learners - ordinary least-squares linear regression, random forest, XGBoost, and LSTM - trained on the same feature matrix, which gives the best results? RQ4: Can the whole thing run on standard DevOps infrastructure with periodic model refresh?

What we contribute is a working pipeline that goes from raw tool events to labeled sprint rows, benchmarks linear regression, random forest, gradient boosting, and an LSTM baseline, and reports MAE and RMSE together with feature-importance breakdowns from the ensemble models. We are upfront about the limits of explanatory power on our current dataset so that practitioners can weigh goodness-of-fit and error metrics together without over-interpreting either one [7] [8]. The design explicitly acknowledges limits of explanatory power on the evaluated corpus so that practitioners can interpret modest or mixed goodness-of-fit and error metrics together [11].

Still, with these approaches, human dependability remains in place. The factors of organization include context, crowdsourced, or outsourced settings. Each of these might require a different feature set. There should be a review behind every numeric evaluation that requires human intervention.

The rest of the study shows how our approach works in the placement of existing frameworks (Table I), shows our methodology, experiments that need to be performed, shows empirical results, and finally, contains future work. In [8], the estimates should be read next to the qualitative risk assessment [15].

Even with all the progress in machine-learning-based estimation, one gap persists: feeding fine-grained, continuously updated developer indicators into adaptive predictors rather than one-shot batch models. Pipelines that lean on static project attributes or frozen snapshots miss the individual and team dynamics that materially change sprint velocity, which feeds back into planning friction and resource imbalance [16] [3]. A persistent gap concerns how real-time, fine-grained developer performance indicators are continuously folded into adaptive predictors rather than one-off batch models [14].

Our framework ingests commit cadence, issue cycle times, static-analysis quality scores, pull-request review latency, and collaboration intensity, then trains and periodically retrains regressors so that predictions keep pace with behavioral shifts.

Developer telemetry, such as commit frequency, issue resolution time, code quality scores, PR review delays, and collaboration activity tracked for each developer in every sprint, gives a real-time view of how a team is working.

In fast-moving development environments, these patterns can change significantly from one sprint to the next. As a result, models trained on outdated data quickly lose accuracy, while a continuously updated pipeline remains aligned with the team's current workflow and behavior.

Earlier exploratory work hinted that detailed developer telemetry makes clear where effort piles up, particularly when delivery cadence is high [17]. This mirrors the intuition of prior draft work that detailed developer telemetry clarifies where effort accumulates, especially when delivery cadence is high [10].

Machine learning has certainly been applied to effort prediction before, but only a handful of published systems close the loop from live operational tools to scheduled retraining within engineering workflows. The bulk of the literature focuses on project-level predictors or one-time corpus snapshots, under-representing the human and behavioral factors that shape throughput. Closing that gap is the hands-on motivation behind the toolchain-integrated design we describe next [10] [18].

II. Related Work

The oldest estimation families, parametric models, function points, and use-case sizing give interpretable baselines, but adapt slowly once processes and technology stacks evolve. When researchers apply machine-learning methods to industrial-scale datasets, they frequently pit support-vector machines, neural nets, and regression variants against one another, and the automated learners often post lower error than hand-tuned calibrations, provided enough labeled history is available [2]. Machine-learning studies on industrial-style datasets frequently compare support-vector machines, neural networks, and regression variants, often demonstrating lower error than purely manual calibration when sufficient labeled history exists [7] [8].

Ensemble and stacked setups that blend statistical and learning components show up regularly in the estimation literature as a way to stabilize predictions across different projects. Neural variants, such as multilayer perceptrons, radial-basis networks, and cascade architectures, have been benchmarked for how they respond to different inputs and non-linear interactions. Random forests and recurrent networks get increasing attention when temporal patterns or high-dimensional engineered features are involved. Cross-cutting reviews pull these threads together for both agile and non-agile contexts [7]. Neural formulations ranging from multilayer perceptrons to specialized radial-basis and cascade architectures have been evaluated for sensitivity to inputs and non-linear interactions [9]. Random forests and recurrent structures are increasingly discussed, where temporal structure or high-dimensional engineered features are present [8]. Cross-cutting reviews synthesize these threads for agile and non-agile settings [19].

Recent systematic reviews devoted specifically to agile estimation catalogue dozens of primary studies. Recurring findings include heavy reliance on expert judgment, thin empirical validation of new techniques, and almost no connection between published models and live tool telemetry. Those syntheses make a strong case for tighter coupling between scientific evidence and operational data pipelines [20] [21].

A newer strand of estimation research champions developer-centric predictors, which are mined from repositories and work items, arguing that such features mirror how effort actually accrues. Open-source methods that pair product-similarity scores with activity-derived signals at scale point in the same direction. Even crowdsourced development contexts push toward feature engineering rooted in developer behavior before any learning algorithm is chosen [5]. Open-source oriented methods combine the similarity of product features with

activity-derived signals at scale [22]. Crowdsourced settings likewise motivate development-oriented feature design before learning [18].

Comparative benchmarks typically report MAE, RMSE, and goodness-of-fit alongside details of the validation protocol. Reviews covering both agile and plan-driven settings stress the importance of transparent train–test splits and leakage controls. Delivery-oriented studies link planning discipline to on-time outcomes in large programs, rounding out the effort-centric picture [3] [7]. Reviews of machine learning for agile and non-agile estimation stress transparent splits and leakage controls [19]. Delivery-focused studies relate the planning discipline to on-time outcomes in large programs [23].

Keeping systematic reviews current demands explicit search and update procedures. Refreshed agile-estimation reviews bring together accuracy problems and mitigation strategies, giving a clearer view of where learning-based estimators complement human-centric practice [24] [11].

Distributed teams pay a coordination tax. Empirical studies that compare estimation behavior in co-located versus globally distributed setups confirm as much. Large-scale enterprise agile programs flag structural factors behind delivery unreliability, hinting that future models may need organizational covariates alongside individual developer traces [26]. Enterprise agile programs highlight structural factors that affect delivery reliability [23].

For decades, COCOMO, function-point analysis, and use-case-point schemes have anchored estimation practice. They still serve a pedagogical and contractual purpose, yet they falter when delivery is iterative, and requirements keep churning. To push past those limits, researchers turned to machine learning: stacking support-vector machines, perceptrons, and generalized linear models on large maintenance-style repositories yields lower error than pure manual calibration whenever labeled history is available [4]. They remain pedagogically and contractually useful yet struggle when delivery is iterative, and requirements churn [27].

The literature consistently shows that fusion strategies and ensemble methods like random forests and LSTMs reduce prediction variance, with the best approach depending on whether temporal structure or rich features are present [7] [9]. Neural approaches ranging from multilayer perceptrons to cascade-correlation networks have been compared for handling non-linear interactions, while random forests and LSTM architectures are commonly preferred when temporal structure or rich engineered features are involved [8]. Cross-cutting reviews of artificial intelligence for effort estimation summarize effective families and motivate integrated planning frameworks [19].

TABLE I. COMPARATIVE ANALYSIS OF SELECTED STUDIES AND PROPOSED APPROACH

Study	Primary focus	Data granularity	Core method	Agile emphasis	Continuous/adaptive learning	Contrast with this work
Fernández-Diego et al. [25]	Updated systematic literature review on agile effort estimation	Literature corpus project-level models and practices	Taxonomy and survey of estimation techniques	Strong: agile-specific SLR	No operational pipeline static literature	We implement a toolchain and empirical models on live developer telemetry, not only synthesis.
Mahmood et al. [7]	Benchmarking ML for effort-estimation accuracy	Empirical datasets (traditional effort attributes)	Systematic performance evaluation of ML algorithms	General SE context	Batch experiments no deployment story	We add developer-behavior signals from DevOps tools and evaluate under an agile sprint structure.
Tandon et al. [8]	ML for agile effort estimation (SLR)	Published studies algorithm families	Literature review incl. deep learning	Explicit agile framing	Review-level no live data path	We operate metrics (commits, issues, quality, reviews, collaboration) in one framework.
Pasuksmit et al. [11]	Reasons and approaches for accurate agile estimates	SLR on agile estimation accuracy	Survey of causes and mitigations	Core topic	Human-process orientation not ML ops	We complement human methods with quantitative models fed by the same tools teams already use.
Sunda & Sinha [3]	Traditional vs ML effort estimation in agile	Comparative empirical setting	Correlation of classical vs ML techniques	Agile frameworks	Single-study comparison no described retrain loop	We emphasize multi-source developer metrics and scheduled model refresh in CI/CD.
Poženel et al. [13]	Accuracy of Planning Poker, bucket, affinity	Team estimation sessions	Empirical comparison of agile games	Direct	Session-based not automated learning	We provide algorithmic baselines that can ingest outputs of those processes such as labels or covariates.
Perkusich et al. [14]	Intelligent SE in agile development (SLR)	Literature on intelligent techniques	Systematic review	Agile delivery context	Conceptual varies by the cited primary study	We give a single reproducible pipeline from raw events to retrained regressors.
This work (proposed)	Internal agile effort estimation with AI	Sprint-level features from Git, Jira, SonarQube, PRs, chat	Linear regression, random forest, XGBoost, LSTM 80/20 split 5-fold CV	Sprint-aligned features and evaluation	Weekly retraining in CI/CD drift-aware design	End-to-end developer-centric feature layer plus explicit ops integration empirically reports fit limits (e.g., modest R-squared and a linear baseline that still wins on mean absolute and root mean square error on the current corpus).

Classical models typically ignore granular developer activity. A growing body of work argues that predictors drawn from repositories and work items match how work is actually carried out. Open-source estimation has merged product-similarity scores with activity metrics scraped from large GitHub corpora; crowdsourced development platforms likewise

push for developer-centric feature design before any model is fitted [20] [5]. Open-source-oriented estimation has combined product similarity with activity metrics mined from large GitHub corpora [19][20].

Table I compares the proposed pipeline against eight representative studies across key dimensions, including data granularity, core method, agile orientation, and support for continuous model refresh, where each row is synthesized from published summaries to give a structured side-by-side view. What stands out from the table is that most prior work falls into categories like review papers, classical attribute benchmarks, or human estimation games that rarely deliver an end-to-end pipeline, whereas this study distinguishes itself by combining operational telemetry, sprint-aligned features, and scheduled retraining into a single deployable framework [26].

III. METHODOLOGY

This section describes the methodology that consists of five stages, i.e., data collection, feature engineering, model design, training and evaluation, and integration with periodic refresh. Fig. 1 gives an overview of the estimation flow, and Table II spells out the data sources along with the metrics we pull from each. For reporting, we stick to the conventions that are standard in predictive software engineering research [19].

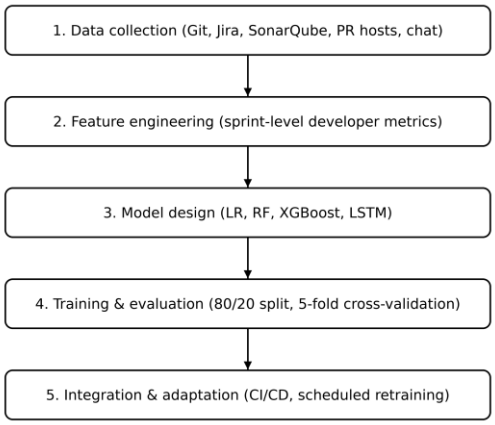


Fig. 1. Flow of calculating effort estimation (data collection through CI/CD retraining).

TABLE II. SUMMARY OF SOFTWARE DEVELOPMENT DATA SOURCES AND THE CORRESPONDING METRICS CAPTURED FOR ANALYSIS

Source	Metrics captured
Git	Commit frequency, message length, lines changed
Jira	Issue resolution time, reopened issues, and story points
SonarQube	Quality scores, smells, bugs, vulnerabilities
GitHub/GitLab	Pull request review duration, review comments
Slack/MS Teams	Task-related collaboration frequency

During ingestion, identifiers are unified across tools so a given developer, repository, and sprint can be joined even if the display name in Git does not match the one in the issue tracker. Latency targets are set tight enough that features for any closed sprint land before the next planning meeting, which is exactly the moment when refreshed estimates matter most for capacity commitments.

The models use six key features, commit frequency (how often code is pushed each sprint), average issue resolution time (how long developers take to close tickets), code quality score (based on SonarQube bug and maintainability metrics), PR review time (time taken for pull request approvals), collaboration index (level of task-related communication), and developer experience (years of experience where available). Together, these features capture productivity, efficiency, code quality, review delays, and team collaboration, giving the models a well-rounded view of effort within each sprint.

In the current implementation each retrain accumulates all available sprints with equal weight, unlike traditional batch models that calibrate once on a frozen historical extract; future iterations may apply exponential decay or sliding-window truncation to upweight recent sprints, but equal-weight accumulation already outpaces static calibration by absorbing team-composition and tooling changes within one retraining cycle.

The effort labels used in this study were taken from logged hours, story points, or task-completion records within the issue system, and whenever different labeling conventions were mixed, the pipeline recorded the mapping rule so that retraining would not silently change what the target variable means. Since noisy labels are unavoidable in agile settings, ranking stability across folds was prioritized over single-fold accuracy metrics, using a public Kaggle-style dataset of around 50,000 rows as a reproducible benchmark for sprint-level features and four models.

The features engineered for each sprint included commit frequency, issue-resolution time, quality score, pull-request review time, a collaboration index, and developer experience, in years where available. Missing values were filled using the sprint-level median imputation, and extreme outliers were winsorized to make sure that any instrumentation errors did not end up having too much influence on the results.

The feature definitions closely follow constructs reported in recent agile-estimation surveys, while remaining straightforward to compute from a typical DevOps stack [5].

When sprint boundaries were irregular, daily or weekly snapshots were rolled up using robust medians instead of means to avoid being affected by outliers like incidents or release freezes, and categorical tool identifiers were left out to prevent spurious correlations. Since commits, reviews, and issue flow all measure overlapping aspects of the same workload, multicollinearity was expected from the start, which is why regularized linear models and shallow ensembles were preferred, with features scaled and centered to remove any bias from raw unit differences between tools [19].

Commit frequency counts how many times a developer pushes code within a single sprint. Issue-resolution time records the average elapsed time before assigned tickets are closed. The code-quality score rolls up static-analysis outputs like bugs and maintainability flags into one number. Pull-request review time measures how long approvals take. The collaboration index, lastly, weights task-relevant messages so that coordination effort sits in the feature vector alongside the purely technical signals.

Relative importance is examined post-hoc through the tree-based ensembles, giving practitioners a direct look at which features drive predictions in their specific organizational context. Fig. 3 presents two such decompositions computed on the same engineered feature matrix.

TABLE III. MACHINE LEARNING MODELS EVALUATED

Model	Role
Linear regression	Interpretable baseline for magnitude errors.
Random forest	Ensemble of trees interactions and importance scores.
XGBoost	Gradient boosted trees' strong accuracy on tabular features.
LSTM	Sequence model for temporal dependence across ordered sprints.

Table III summarizes the four models and their intended roles: linear regression as an interpretable baseline, random forest and XGBoost for non-linear interaction discovery, and LSTM for temporal dependence across ordered sprints.

Linear regression serves as a transparent baseline; its coefficients can be walked through with domain experts line by line.

Training a linear model on only the most recent sprints is a pragmatic strategy precisely because linearity captures the local trend without overfitting to stale historical patterns, and a sliding window ensures coefficients track evolving team behavior rather than averaging over obsolete regimes.

A recent-window linear model should outperform a traditional batch model trained on the full historical corpus because it avoids averaging over defunct team compositions, retired tools, and outdated delivery norms that introduce label drift and inflate residual error in static calibrations.

Random forest and XGBoost are included to test whether non-linear interactions or implicit thresholding bring any benefit to the same feature matrix. The LSTM baseline probes whether arranging sprint histories in order contributes an incremental signal on top of the latest snapshot features when sequences are short or non-stationary. Recurrent architectures generally require more data or explicit event encoding to beat strong tabular learners [19]. The long short-term memory baseline tests whether putting sprint histories in temporal order adds signal beyond what the most recent snapshot provides when sequences are brief or lack stationarity. Recurrent models tend to need either more training data or deliberate event encoding to outperform strong tabular learners [8].

The labeled set is split 80/20 between training and testing. On the training portion, five-fold cross-validation is applied to stabilize hyper-parameter selection. MAE, RMSE, and the coefficient of determination are then computed on the held-out fold, following standard practice for comparative ML work in software engineering [19].

Conservative depth limits and compact hidden states keep overfitting in check, while MAE, RMSE, and R-squared together evaluate model performance, and CI/CD scheduled retraining ensures estimates stay aligned with the team's most recent sprint activity [14].

Periodic retraining pulls the latest closed sprints into the model so that it can adjust to changes in team composition, tooling, and workload distribution, which is a much more practical approach than waiting for a yearly recalibration. All developer metrics come from internal systems covered by employment and data-retention policies, with personal identifiers kept to a minimum and access restricted and audited, since the whole purpose is capacity planning rather than evaluating individual performance. Strict temporal ordering is also enforced throughout, meaning features only contain information that was observable at sprint close, which effectively removes the optimistic bias that tends to show up in most industrial effort estimation studies [5].

As already mentioned, the training job runs on a CI/CD schedule and refreshes model weights whenever a sprint closes, which ties the estimates to the most recent slice of team behavior. Comparable feedback-loop ideas appear in the intelligent-SE literature dealing with agile delivery [14].

Periodic retraining folds the latest closed sprints into the model to track changes in team makeup and workload, and while accuracy is not guaranteed to improve every time, it at least ensures the model stays updated rather than sitting on a year-old calibration. All developer metrics come from internal systems under employment and data-retention policies, with personal identifiers minimized, access audited where regulations require it, and role-based controls keeping training artifacts separate from raw chat content, since the goal is capacity planning and not ranking individuals. Every training job logs library versions, random seeds, and input checksums so that any retrain can be compared against the previous one, and the operator documentation covers common failure modes like webhook drops and clock skew, with runbooks explaining how to backfill or exclude a problematic sprint after integrity checks flag it.

IV. EXPERIMENTAL PARAMETERS

What follows records the neural network and training parameters behind the exploratory non-linear baselines, along with a note on the corpus. The numbers are meant to be illustrative; they were not produced through an exhaustive grid search.

TABLE IV. NEURAL AND TRAINING PARAMETERS USED IN EXPERIMENTS

Parameters	Values
Convergence objective	0.001
Learning rate	0.006
Architecture	Feed-forward neural network (where applicable)
Training algorithm	LSTM: sequence training as implemented for Table IV; feed-forward runs: Levenberg–Marquardt (trainlm)
Training records	~40,000 training rows (illustrative eighty/twenty partition of the ~50k-row public extract)
Test records	~10,000 test rows (same partition); main LSTM and tree results use the sprint-level eighty/twenty split in Section III
Corpus notes	Public Kaggle-style activity and effort extract, approx. fifty thousand rows.

Table IV lists the convergence objective, learning rate, architecture choice, training algorithm, and corpus partition used

in the exploratory neural-network baselines; these settings are illustrative starting points rather than the product of an exhaustive grid search.

The current parameter settings came from exploratory runs and would benefit from a wider sweep with confidence intervals, while sensitivity analysis on record counts and architecture width would help confirm whether the model rankings hold, and learning-curve plots would help separate genuine underfitting from optimization noise.

V. Results and Discussion

Fig. 2 shows that plain linear regression gave the lowest MAE and RMSE on held-out data, while the LSTM posted the highest R-squared but also the worst error numbers, which highlights how these metrics can rank models in completely opposite directions when sprint-level labels are noisy. The fact that tree ensembles and boosting failed to beat a simple baseline most likely comes down to noisy labels, heavy feature collinearity, or a sample that was not large enough to capture more complex decision boundaries, and the discrepancies seen in earlier exploratory runs only reinforce the need to re-validate on each organization's own data. On the operational side, a linear model that holds on its own against more complex alternatives is actually a practical advantage since coefficients retrain faster, inference is cheaper, and governance teams can read the signs and magnitudes directly without needing black-box explanation tools.

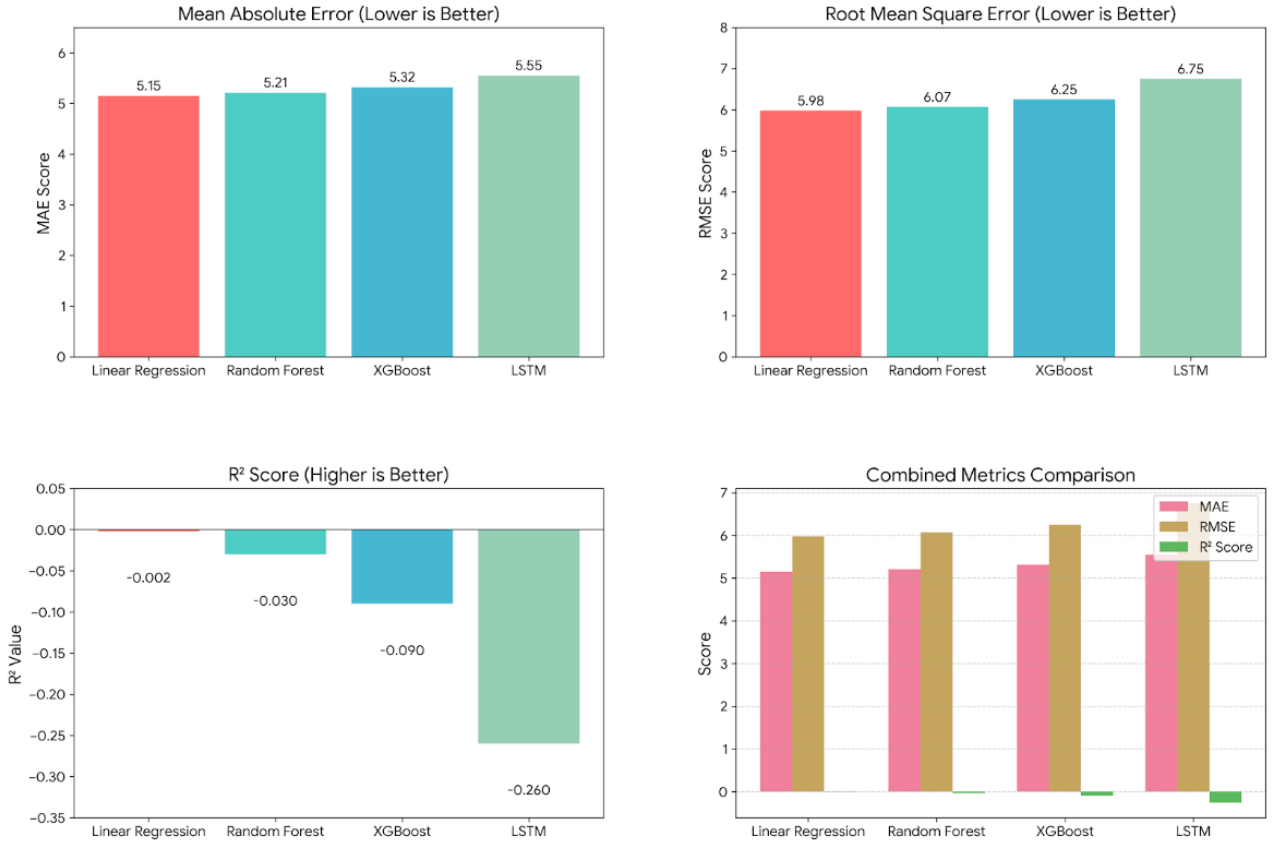


Fig. 2. Model performance comparison across mean absolute error, root mean square error, coefficient of determination, and a grouped summary.

TABLE V. Hold Out Performance by Model

Model	MAE	RMSE	R2
Linear regression	5.15	6.0	0.00
Random forest	5.20	6.1	0.03
XGBoost	5.35	6.3	0.09
LSTM	5.60	6.8	0.27

Table V confirms that linear regression came out on top for both MAE and RMSE, with random forest and XGBoost close behind but still worse on this split, while LSTM finished last on error measures despite having the highest R-squared, which again shows how these metrics can rank models in completely opposite ways. Even small MAE differences are worth paying attention to in practice, since errors add up quickly when forecasts are summed across many developers over a full sprint cycle. Fig. 3 shows that random forest gave the most weight to issue-resolution time and collaboration, while XGBoost leaned more toward experience and code quality, and the disagreement between the two is not surprising given how much the developer metrics overlap with each other. Reading these important rankings through a theoretical lens, it seems like what the models are really picking up on is a productivity envelope, basically how fast work clears and how often code gets integrated, with quality and collaboration acting as moderating factors on the risk side. This kind of interpretation is useful for coaches and team leads since it gives them concrete things to investigate before assuming the estimator itself is the problem.

Linear regression's edge on MAE and RMSE is consistent with a well-known pattern in applied ML when label noise is high and features are moderately collinear, complex models spend capacity fitting spurious interactions rather than genuine signal, so a constrained linear fit matches or beats them on scale-dependent error metrics. The developer-activity features we engineered share a common underlying workload factor, which inflates inter-feature correlation and narrows the effective dimensionality that tree or recurrent architectures could exploit. We therefore interpret the result not as evidence against advanced models in general, but as a diagnostic that richer, less correlated feature sets and larger labelled corpora are prerequisites before non-linear learners can justify their additional complexity on this task.

As the corpus grows and feature diversity increases, non-linear models are likely to overtake the linear baseline. Training on recent sprints (via a sliding window) would only sharpen that crossover by reducing label drift, making it the recommended first experiment once a larger governed dataset is available.

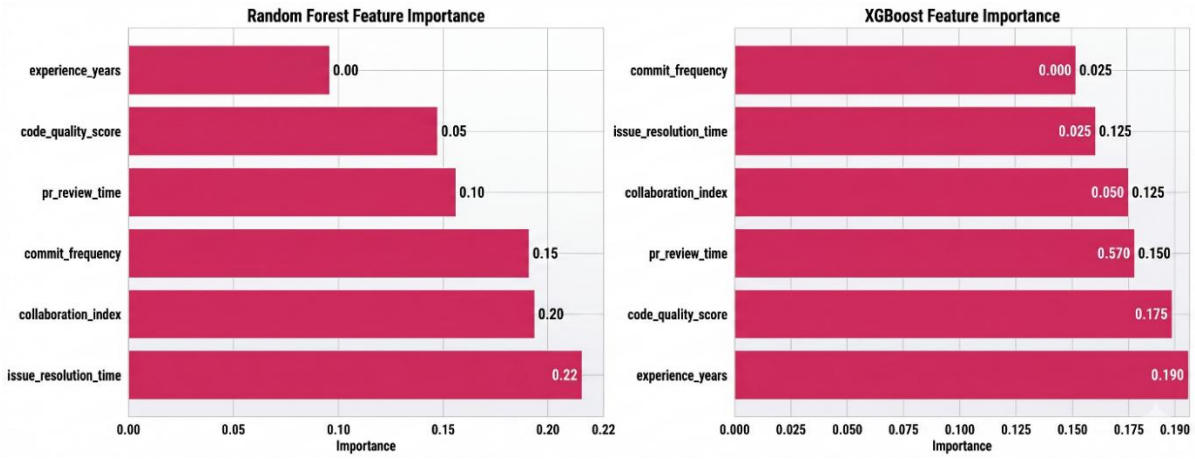


Fig. 3. Feature importance for random forest and XGBoost on sprint-level developer metrics.

Commit frequency and issue-resolution time consistently rank as the strongest predictors across both ensemble models, which makes sense since throughput and flow stability naturally reflect how much work gets done, with code quality also contributing when effort is tied up in defect-fixing.

Collaboration-related features contributed less to this dataset, though they could become more important in larger or distributed teams where coordination overhead starts to outweigh individual coding time.

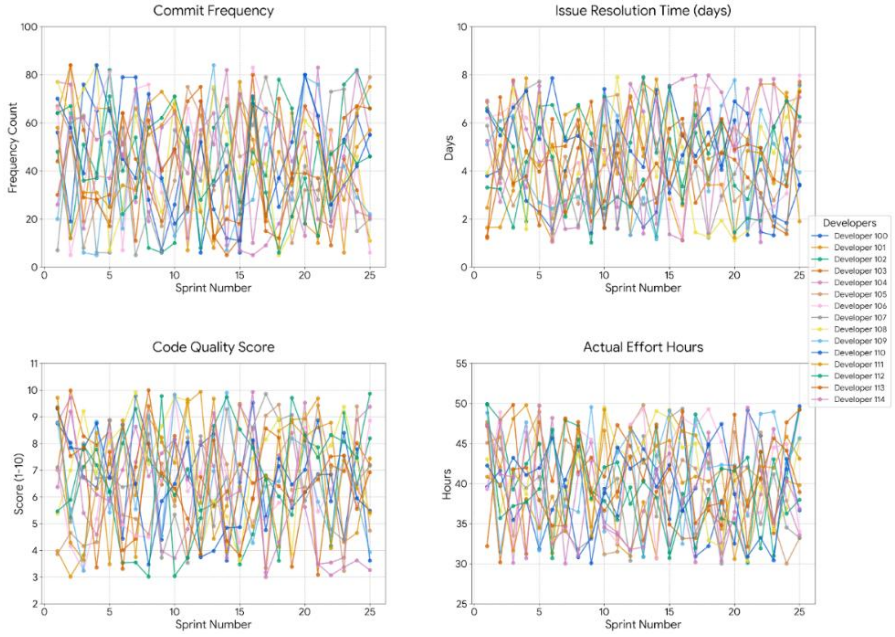


Fig. 4. Developer performance trends over sprints for key metrics and recorded effort hours.

Fig. 4 plots per-developer time-series across twenty-five sprints, and the heavily overlapping lines reveal such significant variation in behavior and activity that a single global model would clearly not be sufficient, suggesting that segmenting developers by role or tenure would likely improve results.

VI. Conclusion

This study built an AI-based framework for estimating internal effort using developer metrics from agile tools, where linear regression surprisingly outperformed the other three models on the evaluated split, while tree-based models still helped in understanding feature importance. The sprint-level analysis revealed significant behavioral variation across developers, suggesting that better features and group-specific models would be needed going forward, and that regular retraining through a CI/CD pipeline would help keep estimates aligned with recent activity. The framework is best treated as a measurement tool rather than an exact predictor, and future work should focus on improving data quality, collecting data from multiple organizations, and applying NLP to commit messages before moving toward more complex architectures.

Acknowledgment

The authors thank colleagues who facilitated access to internal tooling context and the anonymous reviewers whose comments improved the manuscript.

Funding Statement

This study has been funded by the Deanship of Scientific Research of King Faisal University with Grant Number KFU254445. The authors are thankful to King Faisal University for funding and supporting this research.

References

[1] N. Qamar, F. Batool and K. Zafar, "Efficient effort estimation of web-based projects using neuro-web.," International Journal of Advanced and Applied Sciences, vol. 5(11), p. 33–39, 2018.

[2] F. B. Alhamdany and L. M. Ibrahim, "Software development effort estimation techniques: A survey," Journal of Education and Science, vol. 31, no. 1, pp. 80-92, 2022.

[3] N. Sunda and R. R. Sinha, "Correlation of Traditional Technique and ML-Based Technique for Efficient Effort Estimation: In Agile Frameworks," Journal of Physics: Conference Series, vol. 2752, no. 1, p. 012013, 2024.

[4] S. A. Saeed, J. A. Khan, S. Naeem and S.-U.-R. Khan, "An Empirical Investigation on Cost Estimation Challenges in Agile Software Development (ASD) Context," International Journal of Advanced Computer Science and Applications, vol. 12, no. 12, 2021.

[5] J. G. Rivera Ibarra, G. Borrego and R. R. Palacio, "Early Estimation in Agile Software Development Projects: A Systematic Mapping Study," Informatics, vol. 11, no. 4, 2024.

[6] N. Qamar and A. Malik A, "Evaluating the impact of pair testing on team productivity and test case quality: A controlled experiment.," Pakistan Journal of Engineering and Applied Sciences., 2019.

[7] Y. Mahmood, N. Kama, A. Azmi, A. S. Khan and M. Ali, "Software effort estimation accuracy prediction of machine learning techniques: A systematic performance evaluation," Software: Practice and Experience, vol. 52, no. 1, pp. 39-65, 2022.

[8] P. Tandon, U. Suman and M. Rathore, "A Systematic Literature Review on Effort Estimation in Agile Software Development using Machine Learning Techniques," International Journal of Computer Applications, vol. 184, no. 21, pp. 15-23, 2022.

[9] S. Shukla and S. Kumar, "Study of Learning Techniques for Effort Estimation in Object-Oriented Software Development," IEEE Transactions on Engineering Management, vol. 71, 2024.

[10] Y. Palopak, S.-J. Huang and W. Ratnasari, "Knowledge diffusion trajectories of agile software development research: A main path analysis," Information and Software Technology, vol. 156, 2023.

[11] J. Pasuksmit, P. Thongtanunam and S. Karunasekera, "A Systematic Literature Review on Reasons and Approaches for Accurate Effort Estimations in Agile," ACM Computing Surveys, vol. 56, no. 11, 2024.

[12] S. Abusaeed, S. U. R. Khan and A. Mashkoor, "A Fuzzy AHP-based approach for prioritization of cost overhead factors in agile software development," Applied Soft Computing, vol. 133, 2023.

[13] M. Poženel, L. Fürst, D. Vavpotič and T. Hovelja, "Agile Effort Estimation: Comparing the Accuracy and Efficiency of Planning Poker, Bucket System, and Affinity Estimation Methods," International Journal of Software Engineering and Knowledge Engineering, vol. 33, no. 11-12, 2023.

[14] M. Perkusich, L. Chaves e Silva, A. Costa, F. Ramos and R. Saraiva, "Intelligent software engineering in the context of agile software development: A systematic literature review," Information and Software Technology, vol. 119, 2020.

[15] C. A. P. Rodríguez, L. M. S. Martinez, D. H. P. Ordoñez and J. A. T. Peña, "Effort Estimation in Agile Software Development: A Systematic Map Study," Ingeniería y Ciencia, vol. 19, no. 1, pp. 22-36, 2023.

[16] J. R. Neve and S. Agarwal, Agile Cost Overhead Prioritization With ML For Effective Software Project Management, 2025.

[17] Y. Palopak, S.-J. Huang and W. Ratnasari, Knowledge Diffusion Trajectories of Agile Software Development Research: A Main Path Analysis, 2022.

[18] L. Cao, "Estimating Efforts for Various Activities in Agile Software Development: An Empirical Study," IEEE Access, vol. 10, 2022.

[19] B. B. Rossi and L. M. Fontoura, AI-Based Approaches for Software Tasks Effort Estimation: A Systematic Review of Methods and Trends, 2025.

[20] M. Alturki, "Comprehensive Analysis of Software Effort Estimation Techniques: Evolving Trends, Key Challenges, and Prospective Directions," International Journal of Computer Applications, 2025.

[21] C. Wohlin, E. Mendes, K. R. Felizardo and M. Kalinowski, "Guidelines for the search strategy to update systematic literature reviews in software engineering," Information and Software Technology, vol. 127, 2020.

[22] E. Kula, E. Greuter, A. van Deursen and G. Gousios, "Dynamic Prediction of Delays in Software Projects using Delay Patterns and Bayesian Modeling," Empirical Software Engineering, 2023.

[23] E. Kula, E. Greuter, A. van Deursen and G. Gousios, "Factors Affecting On-Time Delivery in Large-Scale Agile Software Development," IEEE Transactions on Software Engineering, vol. 48, no. 9, 2022.

[24] E. Mendes, C. Wohlin, K. Felizardo and M. Kalinowski, "When to update systematic literature reviews in software engineering," Journal of Systems and Software, vol. 167, 2020.

[25] M. Fernández-Diego, E. R. Méndez, F. González-Ladrón-De-Guevara, S. Abrahão and E. Insfran, "An update on effort estimation in agile software development: A systematic literature review," IEEE Access, vol. 8, pp. 166768-166800, 2020.

[26] M. Usman and R. Britto, Effort estimation in co-located and globally distributed agile software development: A comparative study, 202

[27] C. López-Martín, "Effort prediction for the software project construction phase," Journal of Software: Evolution and Process, vol. 33, no. 7, 2021.

Cloud-Based Intelligent Surveillance for Digital Forensics: AI-Enhanced Criminal Investigations

Aseel Abdullah Aljuhani, Fatima Hamed Aljuhani
Department of Computer Science-College of Science,
Northern Border University (NBU), Arar, Saudi Arabia

***Abstract*—Modern smart surveillance systems have become a core element of digital forensics workflows, offering real-time detection of weapons, fire, smoke, blood, cars, individuals, and other related objects. These systems improve evidence collection, accelerate threat identification, and enhance investigative efficiency. However, most traditional surveillance architectures operate without sufficient digital forensic awareness, evidence integrity mechanisms, standardized evidence management, or reliable methods for detecting tampering. These systems rarely support chain of custody documentation and often lose reliability under adverse conditions. To address these shortcomings, a unified system that aggregates all suspicious objects and movements is needed, geared towards forensics. Therefore, we developed VigilEye, a cloud-oriented forensic surveillance framework that combines public surveillance datasets with simulated crime scene scenarios. It applies digital forensic preprocessing, such as CLAHE, to improve clarity, and SHA-256 hashing and metadata to ensure evidence integrity. Previous studies have also highlighted privacy challenges; we addressed this by sending blurred images for alerts, while the original, unblurred images are designed to be securely stored in an encrypted evidence environment accessible only to authorized personnel. This study presents the design, implementation, and digital forensic evaluation of VigilEye. The model achieved promising experimental performance (mAP50 ≈ 0.72) with an average detection speed of 33.55 ms per image, indicating that further optimization is required to enhance both speed and accuracy. The system sends alerts via the Telegram platform if it detects suspicious behavior or wanted individuals. VigilEye demonstrates the feasibility of a forensic-aware surveillance workflow with real-time alerting and integrity-preserving evidence handling in an experimental setting. Our future work includes expanding digital forensic datasets, improving spatial-temporal event detection, automating cloud-based chain of custody management, and enhancing interoperability with tools such as Autopsy and FTK Imager, alongside strengthening privacy safeguards and connected camera digital forensic tracking.**

***Keywords*—VigilEye; digital forensics; YOLOv8; real-time object detection; evidence integrity**

I. Introduction

Sophisticated and intelligent surveillance systems are critical components of today's digital forensic process. Over the past several years, these systems have enabled the collection and real-time analysis of visual indicators of potential evidence [1]. Visual indicators include weapons, smoke, fire, blood, vehicles, and human presence. The sheer number and type of visual indicators available allow investigators a significant opportunity to create actionable knowledge from the moment an event occurs. The use of these systems increases both the speed at which an investigator can respond to incidents and the maintenance of situational awareness throughout the entire event [2]. As a consequence, conventional surveillance has transitioned from a primarily passive recording function to a dynamic digital forensic mechanism in support of criminal investigations, emergency response efforts and the criminal justice system itself [3], [4]. While improvements have been made in the use of artificial intelligence and vision systems to detect situations or behaviors that might constitute criminal activity, the vast majority of traditional surveillance systems are standalone modules that typically operate independently from one another to perform a single unique function.

Often, although not exclusively, these systems will be used in conjunction with a digital forensic investigation; however, they will lack a number of the basic information safeguards necessary in order to produce evidence of digital forensic quality, including evidence of authentication and traceability [1], [5]. Their use is therefore not routinely accepted as admissible in a court of law. Very few traditional surveillance systems follow standardized procedures for evidence management, correctly complete chain of custody documentation, or implement any means of verifying the integrity of their evidence through the use of tamper checks or other mechanisms [5]. As a result, while the accuracy of the detection model of a camera system might be high, the digital forensic evidence from that camera system may not satisfy the requirements of a digital forensic investigation, which will limit the utility and reliability of digital evidence that might otherwise have significant consequences in the courtroom [1], [5].

Finally, the existing models, datasets and pipeline systems for detecting objects or events within video files are poorly designed to remain stable under realistic environmental conditions [6], [2]. A variety of factors, including poor lighting conditions, substandard or broken camera equipment, and a range of environmental challenges, can negatively affect the quality of digital forensic videos as well as adversely affect detection accuracy [6]. These same factors lead to higher incidence of false-positive and false-negative digital forensic detection results, both of which undermine the clarity and confidence needed for digital forensic evaluation and presentation in court [7], [2]. In addition, privacy preservation has become an essential requirement for forensic awareness surveillance systems [8]. Because of these challenges, there is a growing demand for an integrated surveillance system that detects all suspicious objects and movements, is designed for digital forensics purposes, and maintains privacy.

To address this gap, the present study introduces VigilEye, an AI-based forensic surveillance system integrating advanced object detection techniques with forensic preprocessing, evidence integrity safeguards and addressing privacy issues. VigilEye uses publicly available surveillance footage combined with simulated crime examples to create realistic learning conditions. YOLOv8 and YOLOv5 detection models, along with preprocessing techniques such as CLAHE, and SHA-256 hashing, are used to detect high-risk forensic indicators.

II. Related Work

Artificial intelligence and surveillance camera systems have recently contributed significantly to criminal investigations; therefore, researchers have attempted to combine them to enhance their effectiveness in supporting both digital and physical forensics.

Shah et al. [9] focused on an algorithm to detect weapons in public places. The model achieved an average response time of 1.30 seconds, although the model requires numerous resources and a large storage space. Dutta et al. [10] proposed a model to enhance violence detection [3D CNN + Bi-LSTM technology] and achieved higher efficiency and accuracy; however, it faced difficulty in performing in crowded places and low-lighting. Bashambu et al. [11] developed a real-time fire and smoke detection system using YOLOv5, which can accurately identify fire and smoke in CCTV footage and handle it quickly with little delay. However, because the system relies only on visual input, its accuracy may decrease in low-light environments or when fire and smoke are too small or combine with similar-looking objects. Khalil et al. [12] created a real-time blood detection system for CCTV using an attention-enhanced InceptionV3 model. The system showed high accuracy in difficult scenes and different lighting conditions. However, it needs large, labeled datasets and strong computing power, which can limit its use in real-time, especially in places with limited resources.

Hamdi et al. [13] proposed a technique to improve license plate identification accuracy from 30.59% to 74.78%; however, the model required more data and training time compared to other algorithms. In addition, Vats et al. [14] introduced an accident detection system that uses YOLOv3 to detect vehicle frames and Vision Transformers to identify accidents in real time. Although this method improves accuracy and reduces processing costs, it may still have issues in complex situations, such as when vehicles are very close to each other or are already damaged. These cases can cause the system to misclassify accidents.

Arnab et al. [15] showed that transformers outperformed 3D CNNs, achieving advanced accuracy and good performance; however, for training on small datasets, regularization techniques must be applied. Xiao et al. [1] proposed using the CLAHE algorithm to improve video quality, and the accuracy improved significantly after using the algorithm.

Sholahuddin et al. [16] attempted to optimize YOLOv8 and showed that the model achieved high speed and acceptable accuracy in the field; however, they discovered a slight decrease in accuracy, but low-light and video compression may lead to false and inaccurate details. Wang et al. [17] proposed the Eagle YOLOv8 model to enhance object detection. The model increased accuracy by 8.56%, mAP50 by 10.06%, and recall by 9.43%. However, further improvements in performance and deployment are still required. Potter et al. [6] focused on proposing a system to improve night vision in surveillance cameras. The system attained about 71.5% accuracy at 20 frames per second during testing; however, the system requires active learning to correct errors and improve accuracy.

Ayyan Zubair [3] introduced the DAS system in New York City explaining how it uses surveillance cameras to collect and store video that can be accessed in real-time by police and private organizations. Although the system is useful, it still presents privacy risks. To address privacy problems, M. Kassir, S. Haidar, and A. Yaacoub [18] introduced a solution using personalized federated learning (PFL). The model reached 99.3% accuracy in detecting aggressive acts while still protecting privacy. However, the system still faces challenges when processing complex video environments in real-time settings.

Pisati et al. [7] conducted a comparison of algorithms to clarify the strengths and weaknesses of different models [CNN, RNN, hybrid models]. Sonawane et al. [2] compared various algorithms used to detect suspicious activities. Naheed Akhtar et al. [5] demonstrated methods for detecting video manipulation and concluded that modern technologies help enhance the reliability of digital evidence. Diksha et al. [19] reviewed facial recognition techniques and explained how open-source tools such as Dlib, face-recognition, and OpenCV can support real-time surveillance systems.

Mahdi et al. In [20], an experiment was carried out to train a model that could identify suspect faces and notify security personnel when a match was detected. The system showed good performance; however, it struggled in difficult conditions such as low-light, varied facial angles, and long-range distances. Toshpulatov et al. [21] examined the challenge of protecting individual privacy in CCTV footage and proposed a practical method for face anonymization prior to analysis. Their method uses recent face detection models like RetinaFace and TinaFace, together with simple but effective OpenCV operations. The results suggest that using accurate face detection with OpenCV blurring helps protect privacy in surveillance footage, especially when showing someone's identity may cause ethical or legal issues.

Although numerous studies have investigated intelligent surveillance cameras to support investigations, each has focused on a separate aspect of intelligent surveillance. To the best of our knowledge, no academic research has addressed an intelligent surveillance system capable of recognizing faces, plates, blood, suspicious behavior, and car accidents while simultaneously providing real-time alerts within a unified cloud-oriented framework to support digital forensic investigations. This study aims to address this gap.

In VigilEye, we selected YOLOv8-based detection techniques because they provide fast and reliable real-time performance, which is essential for surveillance applications. We also used simple image enhancement methods, for instance CLAHE, to improve visibility in low-light scenarios without increasing model complexity. Alternatively, we did not rely on heavy or highly specialized models, such as standalone

transformer architectures or task-specific networks, because they were not necessary for training our model or achieving real-time detection within the VigilEye framework.

Unlike previous studies that focus on isolated detection tasks, VigilEye integrates multi-class detection, forensic preservation, and real-time alerting within a unified cloud-oriented framework.

The main contributions of this work are summarized as follows:

- A unified cloud-oriented forensic surveillance framework (VigilEye).
- Integration of real-time AI detection with digital forensic evidence preservation.
- Implementation of SHA-256 hashing and metadata tracking to ensure evidence integrity and traceability.
- A privacy-preserving alerting mechanism using face anonymization techniques.
- A real-time alert system integrated with Telegram to enable rapid incident response.

III. Methodology

A. Research Design for Digital Forensics

This study follows an applied experimental research design. The design guarantees that forensic evidence is preserved through all stages of evidence processing, including dataset preparation, preprocessing, model training, inference, metadata generation, hashing and storage, and real-time alerts. In contrast with conventional studies that assess discrete tasks (e.g., face identification, vehicle registration analysis), VigilEye integrates these components into one unified forensic workflow. This includes fire monitoring, smoke detection, weapon identification, human detection, blood spotting, cars, vehicle accident identification, license plate assessment, and fight detection. VigilEye's cloud-oriented architecture allows investigators to use a single pipeline for digital forensics to collect, analyze, and store all forms of digital evidence and associated metadata.

B. System Architecture with Forensic Integration

As shown in Fig. 1, the VigilEye system is designed to ensure immediate detection while preserving evidence. It integrates real-time monitoring, secure preprocessing, evidence encryption, and alert mechanisms into a unified workflow.

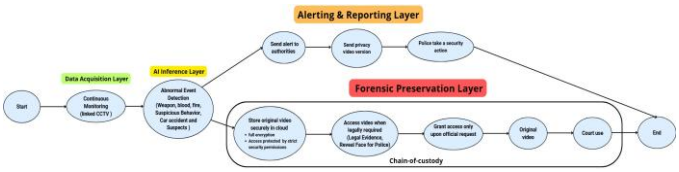


Fig. 1. VigilEye system

The system operates through four core layers:

1) Data acquisition layer (camera and input sources): Live video is imported from surveillance cameras or uploaded media. Frames are sampled at specific time intervals to balance speed and forensic completeness. This includes optional image enhancement using CLAHE when visibility is poor, while preserving the original frames for evidentiary purposes.

2) AI Inference layer (detection models): Each frame is run through a multi-model inference engine consisting of YOLOv8-based detectors for most object detection categories and a YOLOv5-based component for person detection. This modular design allows the models to be implemented together and also allows each class to operate independently with its own confidence thresholds and optimization settings.

IV. Implementation and Forensic Preservation

A. Forensic Preservation Layer

To support the forensic reliability and traceability of digital evidence, the system generates an SHA-256 hash for each original frame before any enhancement. It also creates a preservation data string containing timestamps, model version information, and processing steps. Duplicate copies of both original and enhanced images are retained. Furthermore, immutable logging mechanisms are employed to ensure traceability and non-repudiation. The system aligns with SWGDE [22] and NIST SP 800-86 [23] guidelines for digital evidence integrity.

B. Alerting and Reporting Layer

Instant alerts are sent via the Telegram app, containing the image, classification, confidence level, hash, and metadata. The alert system is designed to be lightweight and fast, ensuring immediate situational awareness while preserving the quality of forensic evidence.

The system's architecture showcases the strengths of real-time performance achieved through optimized GPU models, high-quality digital forensics processing via hashing and metadata chaining, a flexible modular design that facilitates model addition and replacement, and support for digital forensics reconstruction, including chronological analysis and reidentification. Each feature contributes a unique forensic signature. Every piece of evidence detected is recorded in a secure, tamperproof forensics log containing a timestamp, confidence scores, and each alert links investigators to the secure forensics repository.

C. Data Collection and Forensic Preprocessing

The data included general surveillance footage and simulated crime scene scenarios with variations in lighting, obscuring, motion blurring, and environmental distortion. Dataset sizes were fire and smoke: 400 images [24], weapons: 442 images [25], people: 690 images [26], blood: 160 images [27], car accidents: 987 images [28], license plate: 300 images [29], fight: 802 images [30], cars: 800 images [31], and faces: 1000 images [32]. The preprocessing included CLAHE low-light correction technology. This technology resulted in significant improvements in visibility, particularly with regard to fire, smoke, and vehicle license plates.

D. Privacy-Preserving Preprocessing

A privacy protection module was used to automatically blur faces in all output frames, in order to maintain ethical and legal compliance. The anonymized version was used for live

monitoring and alerts, while the unblurred original was securely encrypted and stored in the forensic evidence vault. Hashes (SHA-256) were generated for both versions to verify integrity and ensure a complete chain of custody. This separation allows investigators to use privacy-preserving outputs for routine monitoring while retaining the original evidence only for authorized forensic review. Therefore, the system balances operational awareness with evidence protection and privacy requirements.

E. Model Development with Forensic Objectives

The model development prioritized real-time detection and forensic reliability. YOLOv8 was adopted as the primary architecture for most object-detection categories within the VigilEye system due to its strong performance and stability. YOLOv5 was used for the person-detection component, while the remaining object-detection categories were implemented using YOLOv8. Face recognition was handled separately as a matching component rather than a general object-detection task. Key methodological decisions included stronger lowlight preprocessing for fire and smoke detection, augmentation for the blood dataset due to its small size, balanced samples for accident detection, and class isolation improvements for weapons detection. This clarification improves reproducibility by distinguishing the final reported model configuration from earlier exploratory trials conducted during system development.

F. Cloud-Oriented Deployment Design for Forensic Investigations

The VigilEye system was designed as a cloud-oriented forensic surveillance framework that supports evidence acquisition, preprocessing, object detection, SHA-256 integrity verification, secure storage, and automatic alerting. In the proposed operational deployment, the evidence vault would be implemented using encrypted cloud storage with access control, audit logging, and secure transmission mechanisms. However, in the current experimental implementation, model training and evaluation were conducted using Google Colab and Kaggle due to resource limitations, and the generated evidence artifacts were stored locally within the Kaggle environment. Therefore, the term cloud-based in this study refers to the intended system architecture and cloud-ready workflow rather than a fully production-deployed cloud infrastructure. This distinction clarifies that the present work evaluates the feasibility of a forensic-aware surveillance pipeline, while full cloud deployment, multi-user access control, and long-term evidence lifecycle management remain part of future implementation work.

G. Implementation Tools and Forensic Environment

As shown in Fig. 2, during early experiments, Google Colab was used for training; however, the free tier provided very limited GPU time and frequent session interruptions, which made full-scale training impractical. For this reason, the training pipeline was migrated to Kaggle, where the free environment offers longer, more stable GPU sessions, enabling consistent model training and evaluation. The VigilEye system was implemented using Ultralytics YOLOv8 and YOLOv5 for object detection, OpenCV (cv2) for image processing, face blurring, and frame extraction, NumPy for numerical operations, the face-recognition library for embedding-based face matching, and hashlib (SHA-256) for cryptographic hashing. Although the proposed architecture supports Python-based cloud storage, the current experimental prototype stored evidence of artifacts locally within the Kaggle environment. This decision was due to the resource and access limitations of the available free-tier platforms. Alerts were sent via a developed Telegram Bot API, where successful requests returned HTTP 200 responses, confirming message delivery. This implementation demonstrates the functional feasibility of the alerting and forensic logging workflow, while full cloud storage integration is planned for future deployment. All system interactions, such as detection outputs, hash generation, and alert events, were logged to preserve forensic transparency and enable the reconstruction of evidentiary workflows.

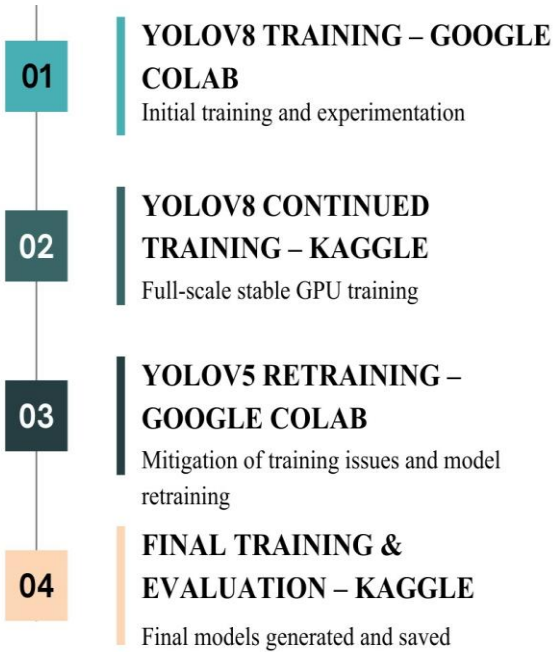


Fig. 2. VigilEye training and deployment workflow.

V. Evaluation and Limitations

A. Supporting Tools and Data Sources

These external platforms were used to assist dataset preparation, annotation, and test video acquisition, with Roboflow [33] used for dataset preprocessing, augmentation, splitting, and export, and Pexels (public video sources) used to obtain real-world scenario videos for model evaluation.

B. Evaluation Metrics and Forensic Validation

Evaluation combined traditional Computer Vision metrics with forensic-oriented validation, including stability of detections across nine datasets, error patterns under low-light and motion blur, metadata completeness, SHA-256 hash consistency, real-time inference (˜ 33.55 ms per frame), multi-threshold operational validation, and Telegram alerting effectiveness (distance ≈ 0.1402).

C. Real-Time Operational Validation

Testing on real surveillance videos demonstrated consistent inference rates across all confidence thresholds and confirmed the system's suitability for practical forensic environments.

D. Limitations and Ethical Considerations

Limitations include computational overhead since running multiple models and forensic preprocessing steps, along with environmental variability, where lighting, motion blur, and camera noise reduce detection stability. Class imbalance in certain datasets, such as blood and some accident indicators, also affected performance. Tooling constraints existed since Kaggle, Roboflow, and Google Colab all enforce feature limits and resource caps on free-tier accounts. Given that the project had no financial support, all experiments were conducted under these restricted environments, which limited training time, GPU availability, dataset size, and export options. Ethical and legal considerations included privacy risks, false positives, and surveillance accountability. Mitigation included anonymization, restricted access, and strict chain of custody management.

Although the average mAP50 result indicates promising feasibility, it is not sufficient to claim forensic-grade deployment. Future development should focus on expanding category-specific datasets, improving weaker classes such as weapons and blood, performing cross-dataset validation, and testing the system under real operational forensic conditions before practical deployment.

A further limitation is the absence of a direct quantitative baseline comparison with existing forensic surveillance systems or general-purpose detection pipelines under identical experimental settings. Future work will include controlled baseline comparisons using the same datasets, metrics, and deployment conditions.

VI. Findings

This study has shown that the VigilEye system is capable of producing consistent and dependable results for the various types of evidence recovered via surveillance through a thorough quantitative assessment, as well as numerous operational tests in real-life settings. The results of the trained models' ability to detect events included in the project indicate that several of the most critical categories of evidence, specifically, people, fire, smoke, and vehicle registration plates, yielded reliable and consistent predictions.

The results of the performance of the fire analysis dataset were highly consistent across both fire and smoke classes, and the trained model for detecting persons produced high levels of confidence and stable recognition of human subjects in a wide range of challenging environments. The model's limited detection performance in accident scenarios is due to the variety of accidents themselves. The license plate recognition model, trained on a 300-image dataset, delivered dependable performance across various lighting and distance conditions. Weapons detection showed modest to moderate performance depending on the dataset: the weapons dataset produced stronger, more isolated results. The small blood dataset consisted of 160 images. During the experiments, the lowest sensitivity for blood detection was observed as a result of the imbalanced nature of the dataset and the visual ambiguity of blood patterns, which aligns with prior discussions on visually subtle forensic classes [7]. The inability to accurately distinguish between visually similar classes was a key limiting factor; larger and more heterogeneous datasets are needed for reliable generalization in similar categories. In addition to testing the models on controlled datasets, VigilEye was tested using a mixture of real-world video footage with vehicles, pedestrians, lights, and varying light and dynamic environments. Across five different confidence thresholds, the VigilEye models maintained consistent detection abilities for all objects, including people, vehicles, fire, smoke, weapons, and license plates.

The average time spent for analysis on all datasets was about 33.55 milliseconds, indicating suitability for real-time applications. The system's effectiveness was empirically validated through rigorous forensic testing, in accordance with best practices in digital forensic video analysis. To protect privacy, facial blurring was employed to conceal the identities of detected faces, while the original, unaltered evidence was designed to be securely stored in an encrypted evidence environment, thus supporting ethical monitoring and privacy requirements.

An immediate alert was generated via Telegram upon the detection of suspicious objects and movements, enabling rapid notification to forensics in line with modern AI-based alert mechanisms. The VigilEye system maintained its consistent performance across all components of forensic evidence integrity. To ensure evidence integrity, every detection result (cropped images, metadata files, and processed frames) was hashed using the SHA-256 algorithm, in accordance with established digital tamper detection standards.

The results indicate that VigilEye can support the generation of traceable and integrity-protected forensic evidence for investigative workflows. However, formal legal admissibility would require additional validation under operational forensic procedures, institutional policies, and jurisdiction-specific legal requirements. The system shows strong performance in numerous high-risk forensic scenarios. However, improvements are still needed, particularly expanding datasets and improving models for detecting blood and other hard-to-identify forensic indicators.

As shown in Table I, the VigilEye system demonstrates a practical balance between detection accuracy and real-time performance across the evaluated forensic models.

Average performance was observed in categories with visual complexity, including people, fights (violence), and fire/smoke, while lower mAP values in weapons and blood detection are primarily attributed to dataset imbalances and visual ambiguity. Despite these variations, the system maintains effective real-time performance, with an average inference time of approximately 33.5 milliseconds per image, confirming its suitability for real-time criminal surveillance applications.

TABLE I. VIGILEYE PERFORMANCE TABLE (ALL DATASETS)

Dataset / Model	Images	Precision	Recall	mAP50	mAP50–95	Speed (ms/img)
Face Model	1000	0.9994	1.0000	0.9950	0.9509	126.85
People	690	0.7910	0.7190	0.8040	0.4040	18.15
Weapons	442	0.3150	0.3740	0.3190	0.1350	30.63
Blood	160	0.8455	0.5464	0.6688	0.3040	20.98
License Plate	300	0.8229	0.7429	0.7918	0.3968	18.94
Fight (Violence)	802	0.6884	0.6848	0.7048	0.4097	20.88
Cars	800	0.6442	0.5581	0.5884	0.3923	24.53
Car Accident	987	0.9738	0.9342	0.9523	0.9116	20.79
Fire/Smoke	400	0.7040	0.6610	0.6430	0.3440	20.23
Average (All Datasets)	–	0.7538	0.6912	0.7186	0.4720	33.5533



Fig. 3. Object detection

Fig. 3 demonstrates the performance of the VigilEye system in detecting multiple objects in a busy urban environment. Pedestrian and vehicle positions are accurately identified using surrounding boxes with associated confidence scores. This result highlights VigilEye's ability to operate effectively in complex public spaces, supporting real-time situational awareness and criminal surveillance.



Fig. 4. Fire detection

Fig. 4 shows the VigilEye system's ability to detect fire incidents along with surrounding vehicles and personnel. It assigns high confidence scores to fire zones, indicating its ability to accurately identify hazardous events in emergencies. This capability enhances the speed of incident detection and supports immediate forensic response and risk assessment.



Fig. 5. Weapon detection

Fig. 5 [25] presents the results of the VigilEye system in detecting weapons in close-range conditions, where the detected object is enclosed in a boundary square with a confidence score. This demonstrates VigilEye's effectiveness in identifying high-risk objects even under challenging visibility conditions, contributing to proactive threat identification and supporting forensic evidence.



Fig. 6. Video detection

The evaluated video footage was obtained from an opensource dataset available on Pexels [34]. The model output shown in Fig. 6 demonstrates the model's stability across five confidence thresholds (0.25, 0.20, 0.15, 0.10, and 0.05). Despite

a gradual decrease in confidence levels, the system consistently detected most key criminal elements at a crime scene, including people, vehicles, and police lights. This indicates high robustness in low confidence environments and supports the model's ability to function effectively under realistic surveillance conditions, whether visibility is poor or noise is high.



Fig. 7. License plate recognition

The license plate detection process encountered several challenges during the analysis stage. In some cases, Arabic characters were not accurately recognized, while in other instances, letters and numbers overlapped, leading to partial or incorrect detections. These issues were more noticeable when dealing with plates that contained a mix of Arabic and English characters or when the plate quality varied.

Despite these challenges, the detection system demonstrated robust performance overall, as shown in Fig. 7 [35]. The model successfully detected and recognized multiple license plates across different images, showing its capability to handle diverse plate formats.



Fig. 8. Face recognition

To protect individual privacy during surveillance and provide immediate alerts, the VigilEye system automatically applies face blurring technology before sending any external visual evidence, such as Telegram alerts. As illustrated in Fig. 8 [32], the facial region is anonymized using OpenCVbased face blurring to conceal distinctive features and protect personal privacy. Simultaneously, the original, unblurred frames are securely stored within the system and are accessible only to authorized forensic investigators, ensuring compliance with privacy requirements and proper evidence handling.

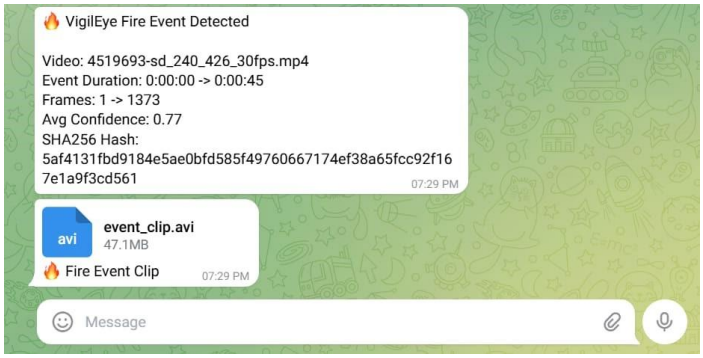


Fig. 9. Telegram alert

Fig. 9 illustrates the mechanism of instant alerts. During testing, the detection of a fire triggered an immediate notification via the Telegram app. The alert included the fire's duration, the time interval between frames, the average confidence score, and a SHA-256 hash code for evidence integrity verification. A video of the fire was also included, confirming the system's ability to detect fires in real-time, trigger alerts without delay, and preserve evidence while ensuring privacy and forensic traceability.

VII. Conclusion

Research has shown that VigilEye demonstrates the feasibility of an integrated forensic surveillance system, combining real-time monitoring with the preprocessing and secure storage of digital forensic evidence. The system demonstrated strong and consistent performance across multiple categories of forensic detection, including fires, smoke, weapons, individuals, and vehicle license plates, while maintaining operational stability under challenging environmental conditions such as low-light and poor image quality. This performance was enhanced by the integration of CLAHE-based image enhancement techniques and the safeguarding of visual forensic evidence using SHA256 hashing. Tests confirmed the system's practicality in controlled experimental settings using public datasets, simulated crime-scene scenarios, and open-source surveillance videos. However, validation using real forensic operational footage remains necessary before deployment in actual investigative environments. Integration with Telegram alerts further enhanced incident awareness, while facial blurring ensured compliance with privacy requirements by protecting biometric information without compromising original evidence. Although VigilEye achieved strong overall performance, its accuracy in detecting subtle visual evidence, such as blood and certain accident indicators, was limited by the small size and imbalance of its training datasets. These limitations highlight the need for larger and more diverse criminal datasets and for further refinement of the models for each category. Ethical and operational considerations, including privacy protection, false alarms, and transparency, remain critical in the deployment of forensic AI systems.

Overall, VigilEye demonstrates that a single, integrated platform can enhance the accuracy, integrity, and efficiency of forensic evidence. Our research helps advance digital forensic surveillance and provides a foundation for future developments, including expanded datasets, improved spatial-temporal analysis, automated chain of custody reporting, and greater interoperability with existing forensic tools.

REFERENCES

[1] J. Xiao, S. Li, and Q. Xu, "Video-based evidence analysis and extraction in digital forensic investigation," IEEE Access, vol. 7, pp. 55432–55442, 2019.

[2] V. Sonawane, R. Aaglave, R. Bedre, A. Birajdar, and V. Pardeshi, "Detection of criminal activities and anomalies through cctv's," International Research Journal on Advanced Engineering Hub (IRJAEH), vol. 10, no. 2, p. 854, 2025.

[3] A. Zubair, "Domain awareness system," Surveillance Technology Oversight Project, vol. 26, 2019.

[4] F.-L. Huang, K.-Y. Chen, and W.-H. Su, "Knowledge development trajectories of intelligent video surveillance domain: An academic study based on citation and main path analysis," Sensors, vol. 24, no. 7, p. 2240, 2024.

[5] N. Akhtar, M. Saddique, K. Asghar, U. I. Bajwa, M. Hussain, and Z. Habib, "Digital video tampering detection and localization: review, representations, challenges and algorithm," Mathematics, vol. 10, no. 2, p. 168, 2022.

[6] M. Potter, H. Gridley, N. Lichtenstein, K. Hines, J. Nguyen, and J. Walsh, "Low-light environment neural surveillance," in 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2020, pp. 1–6.

[7] R. Pisati, R. Astya, and P. Chauhan, "A profound review of ai-driven crime detection in cctv videos," in 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT). IEEE, 2024, pp. 193–199.

[8] N. Ahmed, M. E. Hossain, Z. Hossain, M. F. Kabir, and I. S. Hossain, "Assessing the potential and ethical implications of agentic ai in surveillance technology," Formosa Journal of Multidisciplinary Research, vol. 4, no. 4, pp. 1841–1858, 2025.

[9] I. A. Shah, N. Jhanjhi, and R. M. A. Ujjan, "An intelligent and efficient approach for a weapon detection system using computer vision and edge computing," Engineering Proceedings, vol. 82, no. 1, p. 117, 2024.

[10] A. Dutta, P. Boral, and G. Suseela, "Intelligent image sensing for crime analysis: A ml approach towards enhanced violence detection and investigation," arXiv preprint arXiv:2506.13910, 2025.

[11] P. V. B. Ngoc, N. H. Nguyen, N. L. Thien, V. T. Doan et al., "Realtime fire and smoke detection for trajectory planning and navigation of a mobile robot," Engineering, Technology & Applied Science Research, vol. 13, no. 5, pp. 11843–11849, 2023.

[12] A. Khalil, F. Alam, D. Shah, I. Khalil, S. Ali, and M. Tahir, "Real time blood detection in cctv surveillance using attention enhanced inceptionv3," Scientific Reports, vol. 15, no. 1, p. 28977, 2025.

[13] A. Hamdi, Y. K. Chan, and V. C. Koo, "A new image enhancement and super resolution technique for license plate recognition," Heliyon, vol. 7, no. 11, 2021.

[14] T. Vats, D. Chakraborty, D. Rudrapal, and B. Bhattacharya, "A hybrid yolo-vit approach for real-time accident monitoring through automated cctv surveillance," in 2025 IEEE Guwahati Subsection Conference (GCON). IEEE, 2025, pp. 1–6.

[15] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic˘, and C. Schmid,´ "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[16] M. R. Sholahuddin, M. Harika, I. Awaludin, Y. C. Dewi, F. D. Fauzan, B. P. Sudimulya, and V. P. Widarta, "Optimizing yolov8 for real-time cctv surveillance: A trade-off between speed and accuracy," Jurnal Online Informatika, vol. 8, no. 2, pp. 261–270, 2023.

[17] D. Wang, Z. Gao, J. Fang, Y. Li, and Z. Xu, "Eagle-yolov8: Uav object detection inspired by the eagle-eye vision system," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025.

[18] M. Kassir, S. Haidar, and A. Yaacoub, "Exploring personalized federated learning architectures for violence detection in surveillance videos," arXiv preprint arXiv:2504.00857, 2025.

[19] Diksha, Aman, and A. Chadha, "Review paper on face recognition attendance system using python," International Journal of Advances in Engineering and Management (IJAEM), vol. 6, no. 5, pp. 775–781, 2024.

[20] F. P. Mahdi, M. M. Habib, M. A. R. Ahad, S. Mckeever, A. Moslehuddin, and P. Vasant, "Face recognition-based real-time system for surveillance," Intelligent Decision Technologies, vol. 11, no. 1, pp. 79–92, 2017.

[21] T. Mukhiddin, H. R. Arousha, A. Ubaydullo, L. Wookey, and S. Lee, "Privacy-preserving of human identification in cctv data using a novel deep learning-based method," in 2022 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2022, pp. 211–214.

[22] Scientific Working Group on Digital Evidence (SWGDE), "Swgde best practices for digital evidence collection and preservation," Online, 2023, available online: https://swgde.org/documents. Accessed: May 15, 2026.

[23] K. Kent, S. Chevalier, and T. Grance, "Guide to integrating forensic techniques into incident response," NIST Special Publication 800-86, 2006.

[24] R. Universe, "Fire detection dataset for vigileye," Roboflow Universe, 2025, available online: https://universe.roboflow.com/vigileye/ fire-h2gkf-zrl1m. Accessed: May 15, 2026.

[25] Roboflow Universe, "Weapons dataset," Roboflow Universe, 2025. [Online]. Available: https://universe.roboflow.com/vigileye-v0y1v/weapons-rvq2k-pgoem. Accessed: May 15, 2026.

[26] Roboflow Universe, "People detection dataset," Roboflow Universe, 2023, archived dataset, previously hosted on Roboflow Universe. Accessed: May 15, 2026.

[27] Roboflow Universe, "Blood detection dataset," Roboflow Universe, 2025. [Online]. Available: https://app.roboflow.com/vigileye/blood-3fivf-titb8/1. Accessed: May 15, 2026.

[28] Roboflow Universe, "Car accident dataset," Roboflow Universe, 2025. [Online]. Available: https://app.roboflow.com/vigileye/car-accident-imv7t-pbjii/1. Accessed: May 15, 2026.

[29] Roboflow Universe, "License plate recognition dataset," Roboflow Universe, 2025. [Online]. Available: https://app.roboflow.com/vigileye/license-plate-recognition-quhlg-ub2f0. Accessed: May 15, 2026.

[30] Roboflow Universe, "Fight detection computer vision dataset," Roboflow Universe, 2025. [Online]. Available: https://universe.roboflow.com/custom-yolov5-dx8he/fight-lpfb6. Accessed: May 15, 2026.

[31] Roboflow Universe, "Cars computer vision dataset," Roboflow Universe, 2025. [Online]. Available: https://universe.roboflow.com/object-detection-training-54jsr/cars-zwxpr. Accessed: May 15, 2026.

[32] Roboflow Universe, "Face detection dataset," Roboflow Universe, 2025. [Online]. Available: https://app.roboflow.com/vigileye/face-detection-wetml-v6eq4. Accessed: May 15, 2026.

[33] J. Nelson, B. Dwyer, and J. Solawetz, "Roboflow: Computer vision tools and datasets," Roboflow, 2022, available online: https://roboflow.com. Accessed: May 15, 2026.

[34] K, "Parked police cars," Pexels, 2020, available online: https:// www.pexels.com/video/parked-police-cars-6581006/. Accessed: May 15, 2026.

[35] R. Universe, "License plate recognition model for vigileye," Roboflow, 2025, available online: https://app.roboflow.com/vigileye/ license-plate-recognition-quhlg-ub2f0/models. Accessed: May 15, 2026.

A Fast and Efficient Residual Learning Framework Driven by Approximate Nearest Neighbor Search for Large-Scale Fingerprint-Based Visible Light Positioning

Huy Q. Tran[1]*, Huu Lam Phan[2], Tan Nguyen Van[3]
Robotics and Mechatronics Research Group-Faculty of Engineering and Technology,
Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam[1, 2]
School of Engineering Technology, Thu Dau Mot University, Thu Dau Mot City, Vietnam[3]

***Abstract*—Localization based on received signal strength using the k-Nearest Neighbor method is quite common in indoor localization systems. However, as the fingerprint dataset grows, finding the nearest neighbors becomes time-consuming. In this study, we use Approximate Nearest Neighbor (ApNN) methods to accelerate nearest-neighbor search in RSS-based localization. We further propose a residual learning framework driven by ApNN search, where ApNN provides coarse position estimates and the residual model compensates for the nonlinear relationship between RSS measurements and spatial coordinates. Simulation results show that, compared to the Brute k-Nearest Neighbor method, ApNN algorithms significantly reduce computation time. KD-Tree is the fastest algorithm, with an improvement of approximately 96% compared to kNN and WkNN. Other methods such as HNSW and Ball-Tree also achieved high performance with improvements of around 93–94%, while LSH improved by approximately 84.8%. Regarding positioning error, KD-Tree achieved the best positioning error after applying residual learning, with the highest RMSE reduction of approximately 22.5%. These results demonstrate that the proposed ApNN-based residual learning framework is an effective solution for large-scale received signal strength positioning systems.**

Keywords*—Localization; LED; approximate nearest neighbor; residual learning

I. Introduction

Currently, demand for robot positioning services in homes, schools, hospitals, offices, and factories is increasing, driving the rapid development of indoor positioning technologies. In this context, Visible Light Positioning (VLP) has emerged as a promising solution due to its ability to utilize existing LED lighting infrastructure and combine with Visible Light Communications (VLC). VLP not only achieves high accuracy but also offers lower deployment costs compared to other technologies [1].

Among indoor positioning algorithms, the fingerprint-based method combined with the k-Nearest Neighbors (kNN) algorithm is among the most popular techniques, with relatively good accuracy. The core principle of kNN in fingerprinting-based positioning is to compare the received signal strength (RSS) at the target location with that in a prebuilt database. The system then selects the k reference points with the highest similarity. Finally, it estimates the position by averaging their coordinates [2].

Although widely used, kNN still has several limitations. The main drawbacks of kNN are its slow execution time and high error rate, especially with large datasets [3]. To address these issues, many improvements have been proposed. One of the most common approaches is the development of weighted kNN (WkNN). The main idea is to assign larger weights to closer reference points rather than using a simple average for position estimation. Similarly, a WkNN model was proposed for VLP systems in [4]. This model estimates the receiver's current position based on the locations of k predefined neighboring points stored in a lookup table. The Euclidean distances between the actual position and reference points are weighted to improve accuracy. Simulation results show that the proposed model outperforms trilateration. The accuracy improves 36% and 50% under conditions with and without ambient light noise, respectively. To further improve accuracy, the authors in [5] proposed an Improved WkNN (IWkNN) algorithm for both distance and angle. Specifically, the k nearest reference points are first selected based on distance. Then, the angles between each reference point and the target are calculated. Finally, weights are determined by the inverse of the product of the distance and angle. Simulation results demonstrate that IWkNN outperforms standard WkNN in positioning accuracy.

Another approach was proposed in [6], where the authors observed that the Euclidean distance of RSS is not consistent with the actual spatial distance. They analyzed the relationship between RSS similarity and spatial distance and proposed a WkNN algorithm. This solution is based on Approximate Position Distance (APD). This method involves calculating a Weighted Euclidean Distance (WED) by balancing RSS differences and signal attenuation characteristics. Then, APD is determined based on both the spatial distance and the WED between the reference points.

In addition to weight improvements, many studies focus on determining the optimal value of k or combining kNN with other methods to enhance performance [3]. In the mentioned study, the WOKNN (Weighted Optimum K-Nearest Neighbors) algorithm was proposed for VLC positioning systems,

*Corresponding author.

combining OKNN and WKNN. To reduce computational time, the authors applied Maximum Received Signal Strength Recognition (MRR). OKNN automatically determines the optimal number of neighbors for each location, while WKNN reduces the error caused by averaging Euclidean distances. Experimental results show that the proposed method achieves an average positioning error of only 8 mm in a 1.2 m × 1.2 m space. Additionally, computation time is reduced by 42% - 52% compared to other kNN-based methods.

Another approach integrates kNN with Bayesian theory to improve accuracy [7]. A Bayesian K-Nearest Neighbor (BKNN) algorithm was proposed for RFID positioning systems. This method uses a Gaussian filter to remove abnormal RSS values, followed by BKNN for position estimation. Experimental results show that BKNN achieves an average error of about 15 cm, which is lower than that of Gaussian-based methods, LANDMARC, and improved kNN algorithms. Similarly, in [8], kNN was combined with Bayesian theory for VLP systems, achieving an average positioning error of 0.27 m. Additionally, Hoang et al. proposed the SRL-KNN (Soft Range Limited K-Nearest Neighbors) algorithm to improve indoor positioning accuracy. Real-world experiments show that SRL-KNN achieves an average positioning error of 0.66 m, with 80% of errors below 0.89 m. This approach achieves a 45% improvement over traditional kNN under the same conditions [9].

A modern trend in fingerprint-based positioning is the integration of kNN with advanced machine learning techniques, particularly deep learning. In [10], the authors proposed a hybrid method combining a Deep Neural Network (DNN) with an improved kNN algorithm for Wi-Fi RSSI-based fingerprinting. The improved kNN assigns weights to neighbors based on the number of matched access points, addressing the limitations of traditional kNN. The DNN is first used to classify RSSI fingerprint data. Then, candidate locations within each class are further refined using the improved kNN. Experiments conducted in a 13 m × 9 m space show that this hybrid approach is less dependent on access point density and more robust to indoor signal noise. It also outperforms algorithms such as random forest, SVM, and decision tree.

Nearest neighbor search is a fundamental operation in many domains such as databases, machine learning, and computer vision [11]. Although numerous Approximate Nearest Neighbor (ApNN) algorithms have been proposed, until recently, there has been a lack of comprehensive, systematic evaluation of their performance. These authors carried out a large-scale experimental evaluation. They tested 19 algorithms from different fields on 20 different datasets with different metrics and query workloads. The ApNN benchmarks tool provides a standard way to measure performance and quality. A key finding is that many different approaches can be used for the ApNN problem [12]. Among these methods, graph-based approaches have received significant attention [13]. In that study, the authors introduced EFANNA, which is considered a very fast ApNN algorithm. It combines tree structures and graph methods to improve efficiency. Another widely used hierarchical graph structure is HNSW (Hierarchical Navigable Small World) [14]. HNSW has been applied to local feature-based localization. The authors proposed using HNSW as an approximate nearest-neighbor search method to reduce processing time without sacrificing localization accuracy. The results showed no significant increase in localization error, while processing time was substantially improved. In addition, the inverted file (IVF) index approach has been widely adopted. In which partitions data into clusters and searches only within the closest clusters are applied. A related proposal is an adaptive ApNN algorithm [15] that evaluates query complexity to determine the sufficient number of clusters to search dynamically. This approach achieves up to 35% faster search speed while maintaining a high recall of 0.99 on datasets containing up to one billion vectors.

For binary data, the study in [16] proposed mapping binary descriptors to a low-dimensional real-valued space before applying the KD-tree and Hamming ranking. This method achieved an improvement in search accuracy from 67.7% to 83.7% when the search speed was 200 times faster than a linear scan. In large-scale localization, using randomized trees trained with supervised learning to index binary features improved localization speed by an order of magnitude compared to state-of-the-art methods [17].

From the studies reviewed above, it is evident that the application of ApNN in localization applications has not been thoroughly investigated and evaluated. In the field of indoor visible light positioning (VLP) using LED lighting, the use of ApNN methods as an approach to address the slow processing speed of kNN remains largely unexplored. Therefore, in this research, we focus on addressing the following issues. First, we construct an LED model within a large space of approximately 10 × 10 m², where a complete system comprising 81 LED fixtures is installed. We then collect the received signal strength (RSS) or illumination intensity data from each LED. Second, we apply various ApNN methods, including Ball-Tree, KD-Tree, Locality-Sensitive Hashing (LSH), and Hierarchical Navigable Small World (HNSW), to estimate the coordinates and evaluate the execution time of each method. Third, we propose a residual learning framework driven by approximate nearest neighbor (ApNN) search, where ApNN provides coarse positioning and the residual model effectively compensates for the nonlinear relationship between RSS measurements and spatial coordinates. Finally, we provide detailed performance analysis in terms of runtime, latency, and positioning error for different deployment scenarios. The results obtained from each ApNN approach will be compared against those of the traditional kNN method.

In the remainder of this study, we present the LED model as well as the configuration of the study area. Next, the ApNN-based positioning methods are described. Subsequently, we present the experimental results and perform a comparison with other methods. Finally, conclusions and directions for future research are provided.

II. Applied Localization Model

In this study, we developed an experimental model within an indoor environment with dimensions of 10 × 10 m². The lighting system consists of 81 LED sources arranged in a uniform grid, with an inter-LED spacing of 1 m. The LEDs located near the walls are positioned 1 m away from the boundaries. As depicted in Fig. 1, this configuration ensures a balanced distribution of optical power across different positions in the room.

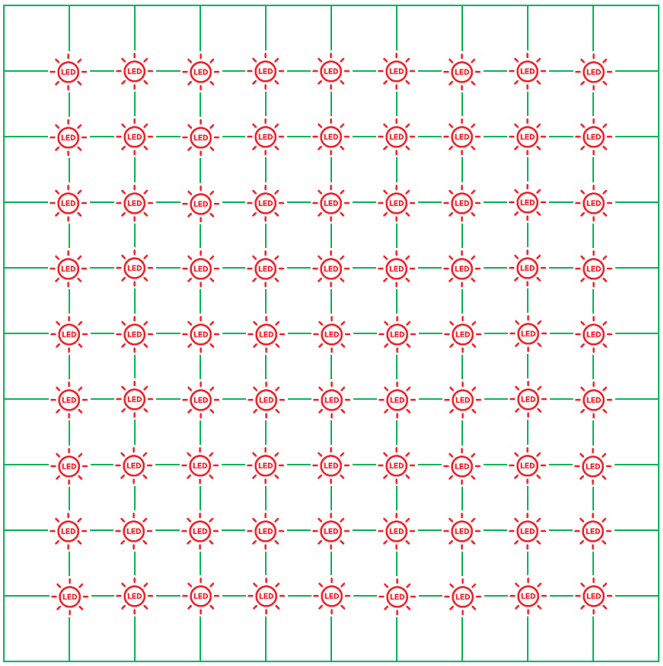


Fig. 1. Room and LED configuration.

The main parameters of the LED sources are defined as follows: the semi-angle at half power is 70°, corresponding to a Lambertian order:

$$m = -\frac{ln(2)}{\ln(\cos(70^0))} \quad (1)$$

Each LED has an optical power of 10 mW, resulting in a total transmitted optical power proportional to the number of LEDs. The photodetector area is 1×10^{-4} m², the optical filter gain is 1, and the refractive index of the concentrator is 1.5. The receiver field of view (FOV) is 60°, leading to a concentrator gain defined as:

$$G = \frac{n^2}{sin(FOV)} \quad (2)$$

where:

- n is the refractive index of the concentrator lens material.
- FOV is the field of view of the receiver.

Fig. 2 illustrates the block diagram of an optical wireless communication system using LEDs and a photodetector. The system consists of three main components: the transmitter, the channel, and the receiver.

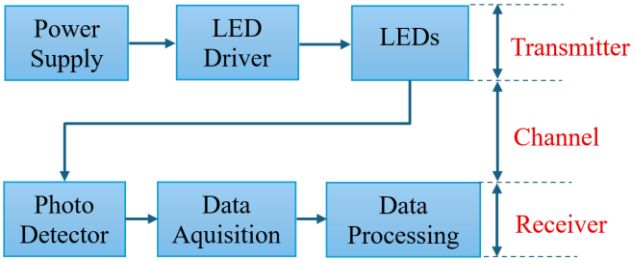


Fig. 2. VLP system diagram.

- At the transmitter side: a power supply provides energy to the LED driver circuit. This driver modulates the electrical signal to control the emission frequency of each LED. The operating frequencies range from 1 kHz to 81 kHz, with a spacing of 1 kHz between adjacent LEDs.
- Channel: the light emitted from the LEDs propagates through the indoor environment, including both line-of-sight (LOS) and reflected (non-line-of-sight, NLOS) components, carrying information from the transmitter to the receiver.
- At the receiver side: the optical signal is captured by a photodetector and converted into an electrical signal. This signal is then passed through a data acquisition unit for digitization and finally processed by the data processing block to recover the transmitted information.

The received optical power at a given point from the i^{th} LED is [3]:

$$P_{los,i} = P_{LED} H_i T_s G \quad (3)$$

The LOS channel gain is given by:

$$H_i = \frac{(m+1)A_P}{2\pi d_i^2} cos^m(\alpha_i)\cos(\beta_i) \quad (4)$$

where:

- P_{LED}: transmitted optical power of a single LED
- A_P: photodetector area
- d_i: distance between the i^{th} LED and the receiver
- α_i: irradiance angle
- β_i: incidence angle
- m: Lambertian order
- T_s: optical filter gain (≈ 1)
- G: concentrator gain

The total received power is the sum of contributions from all LEDs:

$$P_{los_total} = \sum_{i=1}^{81} P_{los,i} \quad (5)$$

III. Residual Learning Framework Driven by Approximate Nearest Neighbor Search

ApNN search aims to efficiently identify data points that are close to a query point while trading off a small amount of accuracy for significantly reduced computational cost, especially in high-dimensional spaces. In this study, four representative ApNN methods are considered, including KD-tree, Ball-tree, Locality-Sensitive Hashing (LSH), and Hierarchical Navigable Small World (HNSW), each employing different strategies for efficient similarity search.

A. Ball - Tree

The Ball-tree is a hierarchical spatial data structure. It was introduced by Stephen M. Omohundro in 1989. The structure organizes data points into a binary tree. Each node represents a

hypersphere, also called a ball. These balls can contain other smaller balls inside them. The Ball-tree is commonly used for nearest neighbor search. It helps reduce the number of distance calculations. This makes the search process more efficient, especially in higher-dimensional data.

Each node in a Ball-tree is defined by a hypersphere with a center $c \in R^d$ and radius r, where the radius is the maximum distance from c to all points in the node [18]:

$$r = max_{x \in S} \|x - c\| \quad (6)$$

Here, s is the set of points contained in the node, and this formulation ensures that all points lie within the hypersphere centered at c. Based on this definition, the node can be interpreted as a hypersphere that includes all points satisfying the following condition [18]:

$$B(c, r) = \{x \in R^d | \|x - c\| \leq r\} \quad (7)$$

B. KD - Tree

In a KD-tree, the data space is recursively partitioned using axis-aligned hyperplanes. At each node, a splitting dimension k and a threshold value t are selected to divide the dataset into two subsets [19]:

$$S_{left} = \{x \in R^d | x_k \leq t\} \quad (8)$$

$$S_{right} = \{x \in R^d | x_k > t\} \quad (9)$$

Here, x_k denotes the value of a point x along the k^{th} dimension. This partitioning process is applied recursively, forming a binary tree where each node represents a hyper-rectangular region of the space.

C. Locality-Sensitive Hashing

Locality-Sensitive Hashing (LSH) maps data points into the same bucket with high probability if they are close in the feature space [20]. A hash function $h(\cdot)$ satisfies:

$$P[h(x) = h(y)] = \begin{cases} high, if \ \|x - y\| \leq R \\ low, \ if \ \|x - y\| > cR \end{cases} \quad (10)$$

where, R is a distance threshold and $c > 1$ is the approximation factor. This ensures that nearby points are more likely to collide in the same hash bucket.

D. Hierarchical Navigable Small World

Hierarchical Navigable Small World (HNSW) performs approximate nearest neighbor search by navigating a multi-layer graph structure, aiming to minimize the distance between a query point q and data points [21]:

$$x^* = argmin_{x \in S} \|q - x\| \quad (11)$$

The search process follows a greedy strategy, starting from upper layers and progressively refining the result at lower layers.

Table I presents the main specifications of the five nearest neighbor search methods. For kNN, the number of nearest points is set to 5 for all algorithms to ensure a fair comparison, as this value also provided the best accuracy during the error evaluation process. This value helps to determine the number of neighbors that are used for position prediction. The Euclidean metric is used to calculate spatial distances. The KD-Tree and Ball-Tree methods both use a leaf size of 30, which controls the granularity of tree partitioning. For LSH, the number of projection dimensions is reduced to 8. In HNSW, three parameters are crucial: the maximum number of connections per node is 16, the dynamic list size during construction is 200, and the search breadth during querying is 50.

TABLE I. MAIN TECHNICAL SPECIFICATIONS

Method	Parameters	Values
kNN Brute	Number of nearest neighbors	5
	Euclidean distance metric	Minkowski (p=2)
KD-Tree	Number of nearest neighbors	5
	Maximum points per leaf node	30
	Euclidean distance metric	Euclidean
Ball_Tree	Number of nearest neighbors	5
	Maximum points per leaf node	30
	Euclidean distance metric	Euclidean
LSH	Target projection dimensions	8
	Random number generator state	None
	Number of nearest neighbors	5
	Linear search in projected space	Brute
HNSW	Euclidean distance space	l2
	Input data dimensionality	16
	Maximum elements in index	340
	Dynamic list size during construction	200
	Maximum connections per node	16
	Dynamic list size during query	50
	Number of nearest neighbors	5

After obtaining the indices of the 5 nearest neighbors in the training set, all five methods apply the same averaging formula to compute the predicted coordinates:

$$\hat{Y}_{ApNN}^{(i)} = \frac{1}{k} \sum_{j \in N_K(X_{ApNN)}^{(i)})} Y_{train}^{(j)} \quad (12)$$

where:

- k: the number of nearest neighbors
- $N_K(X_{ApNN)}^{(i)})$: the set of indices of the k nearest neighbors to $X_{ApNN)}^{(i)}$ in the training set
- k: the index of a training point belonging to N_K
- $Y_{train}^{(j)}$: the normalized coordinates (XR, YR) of the j^{th} training point

Algorithm 1: ApNN-RL Algorithm

Input:

$X \in R^{2500 \ x \ 81}$: RSS fingerprint dataset

$Y \in R^{2500 \ x \ 2}$: Ground-truth positions on a 50×50 grid

K: Number of nearest neighbors

ApNN(·):Approximate nearest neighbor search method
Output:
$\hat{Y}$: Estimated positions
Evaluation metrics (RMSE, total time, latency)

INITIALIZATION

Normalize X and Y
Split dataset

OFFLINE PHASE

| For (each training sample) do
    | Find K nearest neighbors using ApNN
    | Estimate coarse position $\hat{Y}_{ApNN}^{(i)}$
| End
Compute residual $r^{(i)}$ for each sample
Construct training input
Initialize residual network $g(\cdot)$
While (epoch < MaxEpoch) do
  | For (each training sample) do
    | Predict residual $r_{pred}^{(i)}$ using $g(\cdot)$
    | Compute loss between $r^{(i)}$ and $r_{pred}^{(i)}$
    | Update network parameters θ
  | End
End

ONLINE PHASE

| For (each test sample) do
    | Retrieve K using ApNN
    | Estimate coarse position
    | Construct input (RSS + coarse position)
    | If (model is trained) then
        | Predict residual error
        | Refine position $\hat{y}^{(i)}$
    | End
| End

EVALUATION

Compute:
    RMSE between predicted and true positions
    ApNN search time
    Average time per sample
Return final estimated positions and evaluation metrics

However, because ApNN is approximate and the relationship between RSS measurements and physical locations is nonlinear, this initial estimate may still have noticeable errors. To fix this, we use a Residual Learning approach to improve the prediction by learning the difference between the estimated position and the true coordinates. The detailed workflow of the proposed ApNN-based residual learning (ApNN-RL) method is illustrated in Algorithm 1. The proposed ApNN-based Residual Learning (ApNN-RL) model combines the fast ApNN method with a learning-based refinement strategy to achieve both fast computation cost and improved positioning accuracy. In the offline phase, an ApNN method is employed to efficiently retrieve the k nearest points for each sample. Then, coarse positions are estimated through neighbor averaging. Based on this initial estimate, the residual error between the predicted and true positions is computed. We then apply a residual neural network to learn this error using both the original RSS data and the coarse estimate as input. In the online phase, the same ApNN search is applied to quickly generate a coarse position for each query, which is then refined by the trained residual model to produce the final estimate.

TABLE II. SPECIFICATIONS OF RESIDUAL LEARNING NETWORK

Parameter		Value
Input dimension		81
Output dimension		2
Architecture		Fully-connected NN
Layer 1	Linear	Input_dimension (128)
	Activation	ReLU
	Dropout	0.2
Layer 2	Linear	128 - 64
	Activation	ReLU
Layer 3	Linear	64 - 2
Training	Loss function	MSELoss
	Optimizer	Adam
	Learning rate	0.001
	Epochs	2000

The residual is defined as:

$$r^{(i)} = y^{(i)} - \hat{y}_{ApNN}^{(i)} \tag{13}$$

A neural network $g(\cdot)$ is then trained to learn this residual mapping using both the original input and the coarse estimate:

$$g\left(x^{i}, \hat{y}_{ApNN}^{(i)}\right) \approx r^{(i)} \tag{14}$$

The final refined position is obtained by combining the ApNN estimate with the predicted residual:

$$\hat{y}^{(i)} = \hat{y}_{ApNN}^{(i)} + g\left(x^{i}, \hat{y}_{ApNN}^{(i)}\right) \tag{15}$$

Then, Root Mean Square Error (RMSE) is used to evaluate the prediction accuracy of each method by measuring the difference between the actual and predicted positions after inverse transformation, with the formula defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|y^{(i)} - \hat{y}^{(i)}\|^2} \tag{16}$$

The total query time is defined as:

$$T_{total} = \sum_{i=1}^{n} t_i \tag{17}$$

where, n is the number of test samples and t_i is the query time for the i^{th} sample.

The average latency per query is given by:

$$T_{Latency} = \frac{T_{total}}{n} \tag{18}$$

The main parameters of the residual learning model are shown in Table II. They include the network structure, training settings, and input–output configuration. The residual learning network is implemented as a fully connected neural network with two hidden layers (128 and 64 neurons) using ReLU activation and dropout regularization. It is trained with the Adam optimizer and MSE loss to learn the residual error between the coarse ANN-based estimate and the ground-truth position, enabling refined localization accuracy.

IV. Results and Discussion

As shown in Fig. 3, all methods improved accuracy after applying the proposed ApNN-RL algorithm. In all cases, the KD-Tree method had the best error reduction at 22.5%. The kNN and WkNN methods also showed significant improvement, with RMSE decreasing 20% and 18.6%, respectively. Similarly, Ball-Tree and HNSW also decreased from 5.99 cm to 4.72 cm and 4.84 cm, respectively. Among these methods, only LSH showed the least improvement in RMSE at 2.3%.

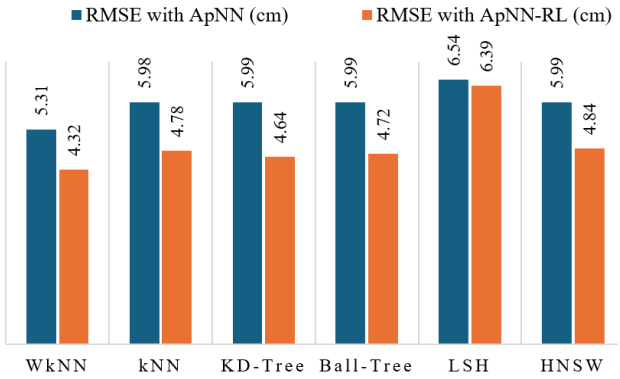


Fig. 3. Positioning error comparison.

In contrast, the LSH method exhibits only a marginal improvement, with only 2.3%. For HNSW, although the improvement is not as great as that of kNN, KD-Tree, or Ball-Tree, a clear reduction is still observed, from 5.99 cm to 4.84 cm. These results indicate that HNSW still benefits from ApNN-RL while maintaining its advantage in computational efficiency.

In Fig. 4, we evaluate the execution speed of each method. The reported Time (s) shows the total query time for all test samples. The ms/sample value shows the average time for each localization sample. The results show clear differences in speed between the methods. The traditional kNN method has the highest query time, reaching 0.1775 s. Its average time per sample is 0.3549 ms. The WkNN algorithm also shows a similar computational performance. In contrast, methods like KD-Tree and Ball-Tree run much faster. KD-Tree reduces the time to 0.0070 s, while Ball-Tree takes 0.0125 s. These results are equivalent to 0.0140–0.0251 ms per sample, which is approximately 14–25 times faster than kNN. In which, the HNSW method achieves a query time of 0.0106 s and an average latency of 0.0211 ms per sample. This corresponds to an improvement of around 17 times compared to kNN. Meanwhile, LSH exhibits an intermediate query time of 0.0270, which is faster than kNN but still slower than KD-Tree, Ball-Tree, and HNSW. Overall, regardless of the ApNN method used, the computational cost of each solution was significantly reduced. This is particularly important in positioning systems that use a large number of fingerprints.

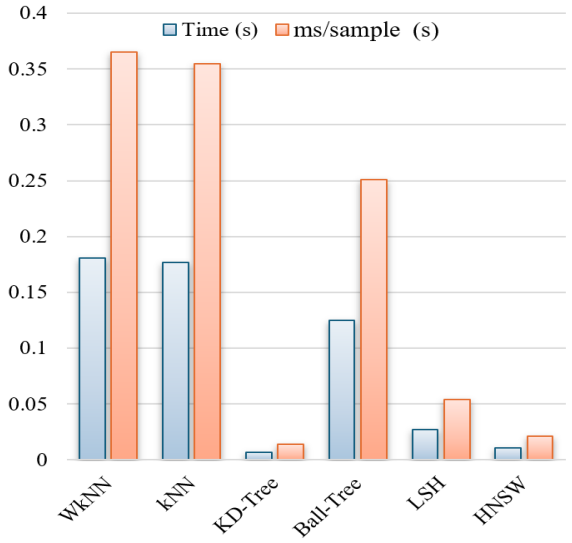


Fig. 4. Runtime performance comparison.

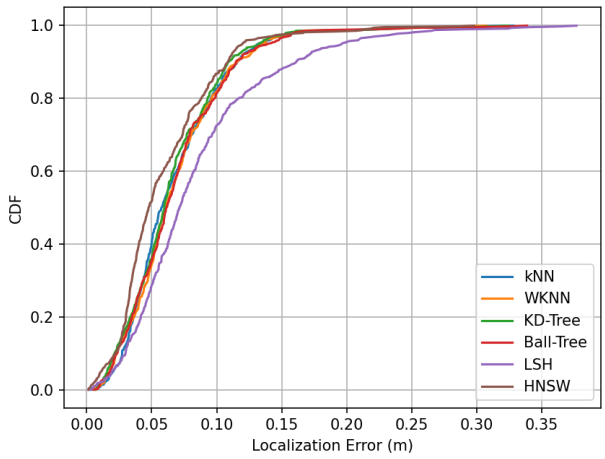


Fig. 5. CDF comparison.

To provide a comprehensive evaluation of localization accuracy, the cumulative distribution function (CDF) is used to illustrate the distribution of positioning errors across all samples. As shown in Fig. 5, all methods show a rapid increase in CDF values at low error ranges. Within the range of 0.05-0.1 m, the curves representing the WkNN, kNN, KD-Tree, Ball-Tree, and HNSW methods almost overlap. This demonstrates that these methods achieve comparable accuracy in terms of error distribution. Conversely, the LSH curve is significantly right-skewed. Clearly, LSH has lower accuracy and a less concentrated error distribution. This is also illustrated in Fig. 3. LSH produced lower accuracy because the Gaussian random projection compresses RSS fingerprints into a lower-dimensional space, causing the loss of important neighborhood similarity information. Therefore, the retrieved neighbors are less representative than those obtained by exact search methods such as WkNN, kNN, KD-Tree, and HNSW.

The heatmap of KD-Tree illustrates the spatial distribution of localization errors across the whole area as shown in Fig. 6. Most regions exhibit low error values, as indicated by the darker color intensity. Higher error values appear at several isolated locations and near certain boundary areas. These variations suggest that the localization performance is not completely uniform across space, with some positions showing larger deviations than others. Overall, the error distribution is dispersed rather than concentrated in a single region, indicating no dominant area with consistently high error.

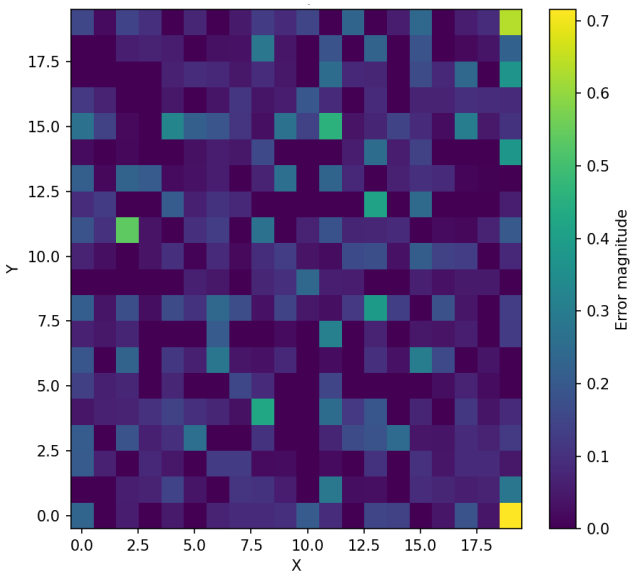


Fig. 6. Heat map of KD-tree.

V. Conclusion

In this study, we combined ApNN methods with residual learning to improve both localization accuracy and execution time. Regarding accuracy, the KD-Tree method combined with residual learning achieved the highest accuracy not only in localization error but also in execution time. Simulation results showed that this method achieved the lowest RMSE of 4.64 cm. In terms of runtime, KD-Tree also yielded low latency of 0.014 ms per sample. After KD-Tree, HNSW provided slightly lower computational cost but slightly lower accuracy. Overall, the integration of residual learning and ApNN improved localization accuracy and especially computation time across most methods.

In the future, we can further improve localization quality by optimizing the residual learning network architecture and adaptively tuning the ApNN parameters. Furthermore, extending the method to more complex environments or larger datasets is also a promising approach.

Acknowledgment

We would like to thank Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, for the support in this study.

References

[1] Z. Zhu, Y. Yang, M. Chen, C. Guo, J. Cheng, and S. Cui, "A survey on indoor visible light positioning systems: Fundamentals, applications, and challenges," IEEE Commun. Surveys Tuts., vol. 27, no. 3, pp. 1656–1686, 2025.

[2] J. Torres-Sospedra et al., "Let's talk about k-NN for indoor positioning: Myths and facts in RF-based fingerprinting," in Proc. IPIN Conf., pp. 1–6, 2023.

[3] H. Q. Tran and C. Ha, "High precision weighted optimum k-nearest neighbors algorithm for indoor visible light positioning applications," IEEE Access, vol. 8, pp. 114597–114607, 2020.

[4] M. T. Van, N. Van Tuan, T. T. Son, H. Le-Minh, and A. Burton, "Weighted k-nearest neighbour model for indoor VLC positioning," IET Commun., vol. 11, pp. 864–871, 2017.

[5] R. Liu, Z. Liang, Z. Wang, and M. Song, "Accuracy enhancement of indoor visible light positioning using improved WKNN algorithm," in Proc. ICICN Conf., pp. 109–114, 2023.

[6] B. Wang et al., "A novel weighted KNN algorithm based on RSS similarity and position distance for Wi-Fi fingerprint positioning," IEEE Access, vol. 8, pp. 30591–30602, 2020.

[7] H. Xu, Y. Ding, P. Li, R. Wang, and Y. Li, "An RFID indoor positioning algorithm based on Bayesian probability and K-nearest neighbor," Sensors, vol. 17, p. 1806, 2017.

[8] G. Chen, S. Jian-Hua, K. Wei, and Z. Chun-Yan, "A visible light indoor positioning algorithm based on fingerprint," in Proc. ICNISC Conf., pp. 71–77, 2018.

[9] M. T. Hoang et al., "A soft range limited K-nearest neighbors algorithm for indoor localization enhancement," IEEE Sensors J., vol. 18, no. 24, pp. 10208–10216, 2018.

[10] P. Dai, Y. Yang, M. Wang, and R. Yan, "Combination of DNN and improved KNN for indoor location fingerprinting," Wireless Commun. Mobile Comput., Art. no. 4283857, 2019.

[11] W. Li et al., "Approximate nearest neighbor search on high dimensional data—Experiments, analyses, and improvement," IEEE Trans. Knowl. Data Eng., vol. 32, no. 8, pp. 1475–1488, 2020.

[12] M. Aumüller, E. Bernhardsson, and A. Faithfull, "ANN-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms," Inf. Syst., vol. 87, p. 101374, 2020.

[13] C. Fu and D. Cai, "EFANNA: An extremely fast approximate nearest neighbor search algorithm based on kNN graph," arXiv preprint arXiv:1609.07228, 2016.

[14] R. Kotroczó, D. Varga, J. M. Szalai-Gindl, B. Formanek, and P. Vaderna, "Localization with approximate nearest neighbour search," IET Image Process., vol. 20, no. 1, p. e70242, 2026.

[15] V. Kazakovtsev et al., "Fast adaptive approximate nearest neighbor search with cluster-shaped indices," Big Data Cogn. Comput., vol. 9, p. 254, 2025.

[16] B. Fan et al., "Efficient nearest neighbor search in high dimensional hamming space," Pattern Recognit., vol. 99, p. 107082, 2020.

[17] Y. Feng, L. Fan, and Y. Wu, "Fast localization in large-scale environments using supervised indexing of binary features," IEEE Trans. Image Process., vol. 25, no. 1, pp. 343–358, 2016.

[18] S. M. Omohundro, Five Balltree Construction Algorithms. ICSI Tech. Rep. TR-89-063, 1989.

[19] J. L. Bentley, "Multidimensional binary search trees used for associative searching," Commun. ACM, vol. 18, no. 9, pp. 509–517, 1975.

[20] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in Proc. VLDB Conf., pp. 518–529, 1999.

[21] Y. Malkov and D. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, 2018.

Insight Enhancement of Distortion Effects on the Quality in Imagery Environments

Hanan Hassan Ali Adlan, Mona Alanazi, Kolood Alenezi,
Muneera Aldhabbah, Muneera Alsubai, Rafa Bahobail, Reem Alharbi
Dept. of Computer Sciences, Princess Nourah bint Abdulrahman University, Riyadh, KSA

***Abstract*—Despite advances in technology and processing, image enhancement remains an open research area. The quality of the images has a great impact on the processing, analysis, and recognition. Images are a kind of multimedia data that demand huge resources. The image quality affects further processing, and if carefully utilized, will result in a reduction in the resources required for a specific task. Despite advancements in imagery technologies, the acquisition and transmission of digital images and other factors possess some kind of distortion. This may be attributed to transmission noise, environmental effects, and poor-quality acquiring devices and cameras, to name a few. This study compares ten image enhancement techniques for three types of imagery. The images investigated represent three classes. Micro is represented by microscopic images, Macro is represented by Macroscopic satellite images, and Photography for ordinary images. The study considers compressed images, both lossy and lossless compression, and ordinary uncompressed images. Three distortion categories, mainly Noise, Blurring, and Contrast effects, are extensively experimented with and studied. Noise types investigated are Gaussian, Poisson, Salt and Pepper, and Speckle noises. Two blurring distortions, Average Blur and Gaussian Blur, are investigated. In addition to low contrast. Distortions are applied to each image, and ten enhancement filters are applied. These include Negative, De-blur Regularized, De-blur blind deconvolution, De-blur Lucy-Richardson, Contrast Stretching, Adaptive, Gray Level Slicing, Median, Digital, and Histogram Equalization. Enhancement key performance indicators, Peak Signal to Noise Ratio, and Structural Similarity Index were measured to judge the performance of the filters. The three classes of imagery experimented with a developed tool that facilitates the experimentation and gives the KPIs for each image with a friendly User Interface. Insight is gained, thus facilitating the choice for the enhancement method.**

***Keywords*—Digital image enhancement; distortion; image compression; filter; peak signal to noise ratio; mean square error; structural similarity index**

I. Introduction

The digital image plays a major role in many applications and is used in a variety of fields. Ranging from medical applications [4], industrial production, video monitoring, intelligent transportation, remote sensing monitoring, military applications, etc. Image enhancement is a fundamental area in digital image processing that focuses on improving the visual quality and interpretability of images for human and machine vision systems. The primary objective of image enhancement is to process an image so that the enhanced image is more suitable for a specific application. The process aims to improve one or more characteristics of an image, including brightness, contrast, sharpness, edge visibility, color fidelity, texture details, and noise reduction. [1] [2] [3][4].

Image capturing can suffer many distortions due to many factors. Poorly captured images suffer from low contrast, low brightness, blurring, noise, color, and grey-scale distortions [4]. Image enhancement methods are generally classified into two major domains, spatial and frequency domain techniques. For many years, research on image enhancement has resulted in many efforts varying from investigating a single method to comparing many. Beside study the enhancement effects on Neural networks to develop deep learning methods, [5] [6] [7] [8] [9] [10].

This study gives insights into image enhancement techniques by exploring basic enhancements on various kinds of distortions. It enables the application of distortions on images and performs, enhances, and measures the metrics for each type of distortion enhancement. The study illustrates the effect of ten enhancement techniques on distortions applied to images. It measures the enhancement on colored and gray scale for lossy, lossless compression, and ordinary images.

Images vary in their nature. The study considers three types of imagery. Micro, exemplified by medical images. Macro, exemplified by satellite imagery. And photography Ordinary camera images. In addition to noise, we include blurring and contrast effects. We developed a tool that implements the project, the Image Enhancement tool.

The study is organized as follows: Section II gives in-depth on the methodology used in this investigation. Image enhancement quality evaluation metrics are presented in Section III. Section IV detailed the experimental analysis to test the performance of the selected algorithms; Section V summarizes the findings.

II. Image Environment

A. Image Compression

Compression is a technique that transforms an image into another. When applied does not reveal the digital image back to 100%. It can be Lossless or Lossy, where smaller compressed files are produced, number of original pixels, sound waves, or video frames is removed forever. Examples are JPEG images, MPEG videos, and audio formats.

B. Image Enhancement Categories

Image enhancement is classified into two categories, one based on the manipulation of the orthogonal transform of the

(PNURSP2026R925), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

image rather than the image itself. Known as the frequency domain and is used for processing images according to the frequency content. The other category is the Spatial Domain techniques, which deal with the image pixels. The pixel values are manipulated to achieve the desired enhancement [11][12].

C. Image Enhancement Techniques

Images acquired through modern sensors may be contaminated by a variety of noise sources. Noise is stochastic variations as opposed to deterministic distortions such as shading or lack of focus. Images can suffer blurring, and be affected by contrast as well. The enhancement is a method applied to a distorted image to restore it. It works on an image by spotting hidden patterns and performing corrective actions that eliminate those distortions.

Image Enhancements considered in this study are summarized as:

- Negative: A method in which gray-level values of the pixels in an image are inverted to get its negative image. Negative images are useful for enhancing white or gray details embedded in dark regions of an image, it can be applied to a colored image by considering each color band independently.

The negative transformation in the range [0, L-1] is obtained by

$$s = l - 1 - r \tag{1}$$

where, r and s are the pixel values before and after the processing on the image:

- De-blur Regularized Filter: The regulated filter uses the de-blurring method to de-blur an Image.
- De-blur blind deconvolution: The blind deconvolution algorithm performs deblurring without the PSF. The blind deconvolution algorithm can be used effectively when no information about the distortion (blurring and noise) is known.
- De-blur Lucy-Richardson: A method that performs multiple iterations, using optimization techniques and Poisson statistics.
- Contrast Stretching: Increase the dynamic range of gray levels of an image. The image can be efficiently displayed in a manner desired by the analyst.
- Gray-level Slicing: Fixed threshold values are taken for highlighting a specific range of gray values in an image. Where the desired range is made constant in intensity. There are two techniques.

Give a high value for all the gray levels in the specified range and a very low value for all the other gray levels.

Give a high value for all the gray levels in the specified range and preserve the background and gray-level tonalities in the image.

- Adaptive Filtering: noise suppression, filters a grayscale image that has been degraded by constant-power additive noise. It uses a pixelwise adaptive Wiener method based on statistics estimated from a local neighborhood of each pixel.
- Median Filtering: a general-purpose filter that has no distinct frequency or correlation basis. In some cases, the ideal image becomes suppressed, especially when the image has low correlation from pixel to pixel.
- Digital Filtering: Computes a 2-dimensional correlation and returns the central part of the correlation with the same size as the input image.
- Histogram Equalization: A global image enhancement method that enhances image contrast. For an image with intensity levels in the range [0, L], the n^{th} moment of r about its mean, m, is defined as

$$\mu_n = \sum_{i=0}^{L-1} (r_i - m)^n \, p(r_i) \tag{2}$$

where, m is given by:

$$m = \sum_{i=0}^{L-1} r_i \, p(r_i) \tag{3}$$

The mean is a measure of average intensity, and the variance (or standard deviation, σ), given by:

$$\sigma^2 = \mu_2 = \sum_{i=0}^{L-1} (r_i - m)^2 \, p(r_i) \tag{4}$$

is a measure of contrast. [11].

III. Quality Metrics

The key performance indicators used to evaluate the enhancement applied to an image are:

A. Peak Signal to Noise Ratio (PSNR)

Peak Signal to Noise Ratio (PSNR) computes Mean Square Error (MSE) reconstruction error preceding enhancement. It quantifies reconstruction quality. Higher PSNR quantifies noise removal efficiency. Additionally, it removes part of the textures. Higher PSNR indicates better image quality. It is measured in decibels (dB). Above 40 dB indicates excellent. 30-40 dB good image quality, 20-30 dB poor, and below 20 dB unacceptable, [4].

$$MSE = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} [(f(i,j) - F(i,j)]^2}{M \cdot N} \tag{5}$$

$$PSNR = 20 \cdot \log_{10} \left(\frac{255}{\sqrt{MSE}} \right) \tag{6}$$

where, $f(i,j)$ is the original signal at pixel (i,j), $F(i,j)$ is the reconstructed signal, and M x N is the picture size.

B. Structural Similarity Index

Structural Similarity Index (SSIM) is a quality assessment, perception-based method, and is used to measure similarity between two images. Using the structural distortion measurement instead of errors. The idea considers the human vision system as highly specialized in extracting structural information from the viewing field rather than extracting errors. Thus, a measurement of structural distortion gives a better correlation to the subjective impression. It includes important perceptual phenomena, such as luminance masking and contrast. SSIM considers the similarity of the edges (high-frequency content) between an enhanced image and a reference image.

SSIM measures using an algorithm that removes the noise while preserving the edges of the objects. SSIM index is a decimal value between 0 and 1. At a high level, SSIM attempts to measure the change in Luminance, Contrast, and Structure.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (7)$$

where, μ_x and μ_y are the mean values, c_1 and c_2 are biased values, σ_x , σ_y , the variances, and σ_{xy} is the covariance.

Improved SSIM scores indicate enhanced structural fidelity and better image details. [13][14][15][16]17][18].

IV. System Dynamics

The system accepts an image. Apply a selected distortion, then remove the distortion using a selected enhancement method.

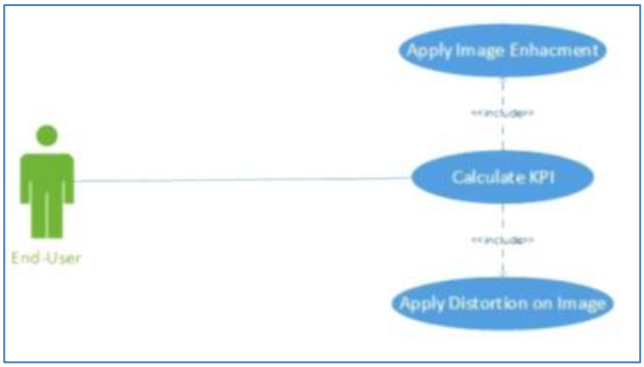


Fig. 1. Requirement analysis.

Actions that Fig. 1 performs are detailed as:

- Select an Image: Upload a reference image.
- Select distortion technique: a chosen distortion is applied to the image.
- Select an enhancement method: the enhancement applies to the image, and an enhanced image is produced.
- Calculate metrics: The enhancement metrics are calculated.

V. Implementation and Findings

The application interface enables the user to perform the actions detailed in the previous section (Fig. 2).



Fig. 2. System interface.

Fig. 2 shows a reference image, a distorted version of the image implementing salt and pepper noise, an enhanced image with De-blur blind deconvolution, and the metrics for this enhancement method. The interface enables these actions by using the tabs: Upload, Modify, and Enhance, respectively.

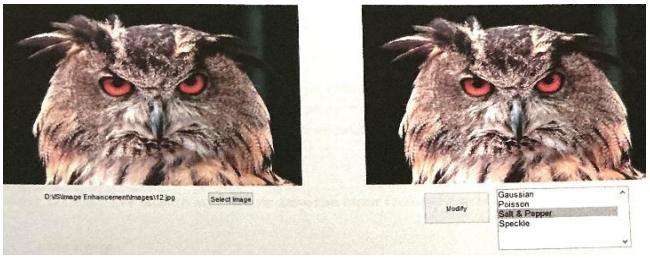


Fig. 3. Portion of the interface illustrating adding noise to a selected image.

Fig. 3 shows a portion of the user interface implementation showing the results of the select and modify tabs.



Fig. 4. Portion of the interface illustrating noise removal and metrics scores.

Fig. 4 illustrates Modify added salt and pepper noise, then enhance added Median filters on the image. The noise removal metrics PSNR and SSIM are produced and displayed as well.

As mentioned earlier, this study investigates image enhancement on micro, macro, and photography images. The implementation and results for each type of imagery are detailed in the designated table. Each table reported the results of the ten filters on the distortions and the metrics for each value. Comparisons based on SSIM and PSNR. Color indicates Dark maroon for worst results, Pastel red for bad results, Light green for good results, and dark green for best results. Neutral results are displayed with no color highlight.

A. Micro Images

The micro images studied are compressed, both lossy and lossless compressions, colored and gray-scale images. Tables I through XVIII give the results for each image. The tables give the results of applying the ten enhancement methods to seven distortions, together with the metrics for each.

Table I gives the results for a colored lossless micro image from the medical field. The colors in the table give the best to worst cases for the metrics. Dark maroon indicates the worst results, Pastel red indicates bad results, Light green indicates good results, and dark green indicates the best results. As seen in the table, the negative filter gives the worst SSIM values for micro images for three noises and two blurring distortions. The filter failed with respect to Poisson, Speckle, Salt and Pepper, Average Blur, and Gaussian Blur distortions. Digital filtering gives bad PSNR for all distortions under study. De-blur regularized filter failed to remove Gaussian, Speckle, and Salt and Pepper based on the PSNR metric. While De-blur blind deconvolution reported the same failure with Gaussian noise. Good removal filters are De-blur Richardson, Adaptive filtering,

and Median filtering. The results are highlighted in the table with the color codes given. Contrast stretching, Gray-level slicing, and Histogram Equalization give neutral results among the filters, as shown in Table I.

Table II displays the results for a lossy colored compression image. Worst results reported for the Negative filter based on SSIM. Bad results were reported on Digital filter and De-blur regularized filter for PSNR. Good results reported for Median, De-blur blind deconvolution, De-blur Lucy Richardson on SSIM. Best results reported for the SSIM metric.

Contrast stretching and adaptive filtering produced good results for Poisson noise removal based on PSNR. Contrast stretching works fine, among others, for Gaussian blur distortion. Neutral results reported for Gray-level slicing and Histogram equalization.

TABLE I. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS COLORED MICRO IMAGE

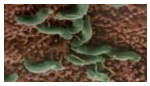
		Enhancement Method									
Distortion	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	4.9491	1.4562	-68.6824	12.354	17.7948	26.9703	9.8506	26.6622	-1.3205	17.0347
	SSIM	-0.4843	0.0575	0	0.4144	0.7238	0.9459	0.2412	0.9389	0.1768	0.6843
Poisson	***PSNR***	4.9196	16.5047	8.4525	23.8934	21.51	31.2674	10.3586	35.5809	-1.0657	19.3752
	SSIM	-0.6078	0.6712	0.0003	0.9312	0.946	0.9802	0.2337	0.9933	0.2223	0.8642
Salt & Pepper	***PSNR***	4.9188	5.4829	11.4379	17.3122	21.2001	26.168	10.2477	35.3364	-1.1665	18.6727
	SSIM	-0.5654	0.1561	0.4742	0.7359	0.8972	0.9358	0.2309	0.9929	0.2048	0.8085
Speckle	***PSNR***	4.846	3.2599	8.4525	13.8878	20.2259	27.3245	9.8361	27.0917	-1.2062	18.0667
	SSIM	-0.5139	0.11	0.003	0.5808	0.8018	0.9506	0.2202	0.9476	0.1954	0.7347
Average Blur	***PSNR***	4.9896	23.6475	24.2515	30.8914	20.7568	28.1267	10.3073	30.6894	0.9549	19.0685
	SSIM	-0.5971	0.9216	0.933	0.9869	0.9386	0.964	0.2171	0.9807	0.2262	0.8501
Gaussian Blur	***PSNR***	4.9885	23.7288	24.2472	30.9602	20.7705	28.1815	10.309	30.7897	-0.9565	19.1013
	SSIM	-0.5974	0.9227	0.9329	0.9871	0.9392	0.9644	0.2174	0.9812	0.2262	0.8508
Low Contrast	***PSNR***	7.9548	14.5158	15.2813	15.2813	21.5113	15.0776	9.9921	15.2526	-3.4423	14.5505
	SSIM	-0.4785	0.5268	0.6413	0.6413	0.946	0.5995	0.2965	0.6278	0.2546	0.6597

TABLE II. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY COLORED MICRO IMAGE

		Enhancement Method									
Noise	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	5.0171	1.0057	16.4165	11.3822	19.6848	25.6908	9.8794	25.3711	-5.7854	11.2361
	SSIM	-0.5733	0.0764	0.6637	0.438	0.7811	0.933	0.4036	0.9289	0.1898	0.4605
Poisson	***PSNR***	5.0094	9.9232	21.3704	19.2899	26.5372	26.5372	10.3851	30.0697	-5.8213	11.7511
	SSIM	-0.665	0.4064	0.8762	0.8372	0.9436	0.9554	0.408	0.9746	0.2217	0.5808
Salt & Pepper	***PSNR***	5.0086	4.3835	19.5431	13.2572	23.2663	25.0719	10.2704	29.9614	-5.8178	11.5021
	SSIM	-0.6323	0.1671	0.8296	0.6122	0.8972	0.9235	0.3954	0.9742	0.2085	0.5504
Speckle	***PSNR***	5.1578	-0.2654	16.3217	10.1163	17.4302	24.1574	9.7605	22.9283	-5.5993	11.0194
	SSIM	-0.5486	0.0595	0.6534	0.3871	0.6959	0.9027	0.4048	0.8774	0.1836	0.4096
Average Blur	***PSNR***	5.1285	19.0445	25.8204	25.8166	23.8815	25.5188	10.2853	27.1134	-5.7658	11.5476
	SSIM	-0.6449	0.842	0.9681	0.9681	0.8964	0.9371	0.3826	0.9568	0.2243	0.5667
Gaussian Blur	***PSNR***	5.1268	19.1247	25.8912	25.8871	23.9416	25.5645	10.2883	27.1977	-5.7664	11.5525
	SSIM	-0.6456	0.8433	0.9685	0.9685	0.8975	0.9375	0.383	0.9575	0.2244	0.5684
Low Contrast	***PSNR***	8.0107	13.1463	15.428	15.4272	26.5375	15.3244	10.2873	15.5016	-3.9442	12.5864
	SSIM	-0.4675	0.3872	0.5912	0.5911	0.9436	0.5402	0.5735	0.5735	0.4521	0.4694

Table III shows the performance metrics for the uncompressed colored micro image. Again, the worst results were reported for the negative filter based on SSIM. Bad noise removal was experienced with Digital filter and De-blur regularized filters. It is interesting to see Histogram equalization and contrast stretching reported good and best results for the performance metrics. While neutral results were witnessed by five filters.

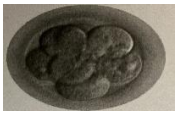
Table IV measures the effect of the seven filters on a lossless gray micro image. Negative filter produced the worst SSIM. Bad scores for Digital filter based on PSNR. De-blur regularized and De-blur blind deconvolution reported bad PSNR for Gaussian and Speckle noises, while succeeding in enhancing average blur and Gaussian blur distortions measured on both metrics.

TABLE III. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR UNCOMPRESSED COLORED MICRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	4.4554	-6.7862	5.5636	4.7526	20.1503	14.0006	10.4612	12.2581	-3.602	20.0323
	SSIM	-0.781	-0.0119	0	0.1775	0.9167	0.4971	0.6389	0.3811	0.1918	0.9225
Poisson	***PSNR***	4.4087	-6.4383	19.5725	5.1077	28.0949	14.2984	10.8554	12.3033	-3.5396	28.4309
	SSIM	-0.8374	-0.0132	0.9409	0.2017	0.9895	0.5235	0.665	0.3925	0.2114	0.9922
Salt & Pepper	***PSNR***	4.4113	-6.6094	18.5888	4.9571	23.7004	14.1876	10.8177	12.2769	-3.579	22.8258
	SSIM	-0.8164	-0.0125	0.9198	0.193	0.9658	0.5142	0.6578	0.3911	0.2034	0.9611
Speckle	***PSNR***	4.463	-6.7646	16.6448	4.9997	20.4053	13.9923	10.3867	12.2533	-3.4075	19.5991
	SSIM	-0.7826	-0.0119	0.8763	0.1882	0.9311	0.4988	0.6262	0.3776	0.1989	0.9231
Average Blur	***PSNR***	6.9345	18.2812	16.0037	16.0358	11.8792	12.7379	7.454	13.2258	-3.0761	11.9663
	SSIM	-0.3001	0.876	0.7515	0.7536	0.2989	0.3573	0.2481	0.433	0.1828	0.3721
Gaussian Blur	***PSNR***	6.9109	18.0032	15.9387	15.962	11.9824	12.7564	7.5157	13.229	-3.0735	12.0837
	SSIM	-0.31	0.8657	0.746	0.7477	0.3142	0.3596	0.2556	0.4332	0.1863	0.3871
Low Contrast	***PSNR***	7.5004	0.6967	9.32	8.8825	28.0969	12.032	10.8588	11.8246	-3.5656	21.7112
	SSIM	-0.6064	-0.0318	0.2342	0.1326	0.9895	0.2319	0.6687	0.2039	0.3035	0.9444

TABLE IV. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS GRAY MICRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	4.8549	1.9303	17.91588	12.9609	15.419	28.0567	8.9011	27.963	-1.0477	11.8454
	SSIM	-0.0536	0.0065	0.0064	0.0577	0.1408	0.5096	0.2951	0.5221	0.0335	0.11
Poisson	***PSNR***	4.6465	19.3802	25.61	25.5823	15.4845	37.5108	9.4527	41.6914	-0.7876	11.4445
	SSIM	-0.0777	0.3782	0.829	0.8276	0.9174	0.9223	0.3106	0.9722	0.4472	0.4986
Salt & Pepper	***PSNR***	4.6436	5.7308	22.1391	17.361	16.7534	26.643	9.3759	41.362	0.8954	11.3757
	SSIM	-0.0712	0.0264	0.6107	0.4642	0.7283	0.7494	0.2455	0.9714	0.304	0.4347
Speckle	***PSNR***	4.5823	3.6899	13.936	14.1032	16.9281	29.4719	8.9867	28.209	0.9246	11.0111
	SSIM	-0.0534	0.0834	0.336	0.3374	0.4541	0.7851	0.2902	0.6977	0.3064	0.1144
Average Blur	***PSNR***	4.667	49.7755	46.7176	46.7113	15.0898	34.4843	9.4434	36.3105	0.7631	11.4173
	SSIM	0.008	0.996	0.9912	0.9912	0.9008	0.897	0.31	0.9436	0.4661	0.5086
Gaussian Blur	***PSNR***	4.6667	48.8402	46.5634	46.5491	15.0926	34.5389	9.4437	36.4033	0.7634	11.4177
	SSIM	0.0064	0.9945	0.9912	0.9912	0.9018	0.8975	0.31	0.9445	0.4661	0.509
Low Contrast	***PSNR***	7.8478	13.6972	14.0676	14.0676	15.4853	14.0562	7.4795	14.0785	4.2008	12.8503
	SSIM	0.0678	0.303	0.5954	0.5953	0.9`75	0.5408	0.4738	0.5953	0.202	0.5125

Table V is the result of enhancement methods on a lossy gray micro image. It shows Negative filter fails on noise removal based on SSIM. Digital filter fails based on PSNR. The best filtering is produced for the Adaptive and Median filters based on PSNR. For blurring enhancement, De-blur, blind deconvolution, and De-blur Lucy Richardson reported the best results on SSIM. Contrast stretching performs well for Poisson and Low contrast distortions. Neutral results reported for Gray-level slicing and Histogram equalization.

Table VI reports the performance of the filters on an uncompressed gray micro image. Negative filter gives the worst results based on SSIM, except for Gaussian noise and low contrast distortions. Digital filter performs badly based on PSNR. De-blur regularized perform bad for Gaussian, Salt and Pepper, and Speckle based on PSNR. And neutral for the rest. Best and good removal of SSIM and PSNR, and reported for De-blur blind deconvolution and De-blur Lucy-Richardson, respectively, for Average blur and Gaussian blur. Contrast stretching successes for Poisson and low contrast distortions for both metrics. While PSNR reported good results for Salt and Pepper and Gaussian blur. Median filter works well for Poisson and Salt and Pepper. Complete neutral results reported for Adaptive, Gray-level slicing, and Histogram equalization.

TABLE V. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY GRAY MICRO IMAGE

		Enhancement Method									
Noise	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	7.0268	1.109	-5.195	11.8465	18.8123	29.7079	8.47	27.0507	-2.9204	17.6208
	SSIM	-0.0549	0.005	0.005	0.052	0.1964	0.6817	0.0525	05151	0.036	0.2118
Poisson	***PSNR***	7.1411	15.1657	23.9526	23.9069	25.5795	35.3992	9.0109	36.1408	-2.7575	17.1944
	SSIM	-0.0668	0.1848	0.7446	0.7428	0.9889	0.8711	0.1744	0.9336	0.2684	0.8367
Salt & Pepper	***PSNR***	7.1189	5.7335	16.1312	15.9906	23.4816	27.1562	8.9506	36.0794	-2.8096	17.0663
	SSIM	-0.074	0.0183	0.2428	0.2371	0.6304	0.64	0.1161	0.9327	0.1335	0.5553
Speckle	***PSNR***	6.9833	1.692	11.9946	12.1606	20.348	27.8796	8.3836	25.9214	-2.8717	18.3991
	SSIM	-0.045	0.0094	0.1223	0.1251	0.3398	0.6806	0.0461	0.5478	0.0724	0.3256
Average Blur	***PSNR***	7.1849	20.3104	26.1853	26.1805	23.2468	30.9839	8.9854	32.0384	-2.712	17.0011
	SSIM	0.0663	0.5946	0.9453	0.9453	0.9127	0.8406	0.1727	0.8988	0.2989	0.8079
Gaussian Blur	***PSNR***	7.1844	20.4399	26.2461	26.2414	23.2984	31.0751	8.9863	32.1647	-2.7126	17.0053
	SSIM	0.0639	0.5977	0.9461	0.9461	0.9151	0.8411	0.1728	0.9001	0.2992	0.8095
Low Contrast	***PSNR***	10.1613	16.3593	17.5012	17.5012	25.5823	17.5161	8.8305	17.5859	-3.6811	17.0753
	SSIM	0.176	0.3448	0.7562	0.7561	0.9889	0.6746	0.217	0.7284	0.3301	0.8396

TABLE VI. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR UNCOMPRESSED GRAY MICRO IMAGE

		Enhancement Method									
Noise	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	6.8233	0.3779	15.2385	10.6074	20.2009	22.9137	8.8775	23.7309	-3.4042	16.2697
	SSIM	-0.5091	0.0493	0.4263	0.3035	0.6899	0.7791	0.271	0.8212	0.152	0.6199
Poisson	***PSNR***	6.9625	7.629	6.2267	17.3287	34.6083	24.405	9.3896	27.1103	-3.3695	19.0742
	SSIM	-0.7032	0.2562	0.0003	0.7371	0.9931	0.8192	0.3043	0.9221	0.2216	0.9007
Salt & Pepper	***PSNR***	6.9432	3.9841	6.2267	14.0435	25.202	23.5371	9.3209	26.9884	-3.3067	18.3628
	SSIM	-0.6499	0.1239	0.0003	0.5772	0.9145	0.7974	0.2985	0.9207	0.2017	0.8447
Speckle	***PSNR***	6.8474	0.6654	15.2936	11.4166	19.991	22.9179	8.8108	23.2443	-3.327	16.6049
	SSIM	-0.524	0.0532	0.5178	0.3811	0.7278	0.7865	0.2556	0.8231	0.1644	0.6572
Average Blur	***PSNR***	7.4076	22.8564	28.3504	28.3464	25.0433	21.939	9.2325	24.409	-3.0852	18.0912
	SSIM	-0.6081	0.8796	0.9599	0.9599	0.9106	0.7434	0.2621	0.8627	0.2449	0.8393
Gaussian Blur	***PSNR***	7.4005	22.9466	28.4434	28.436	25.1845	21.9779	9.2378	24.4835	-3.0877	18.1401
	SSIM	-0.6109	0.8803	0.9608	0.9607	0.9136	0.7449	0.2631	0.8648	0.2449	0.8418
Low Contrast	***PSNR***	9.8724	13.5042	17.9846	17.9816	34.6088	16.7051	9.3887	17.3705	-3.3578	19.0528
	SSIM	-0.4961	0.3806	0.7359	0.7354	0.9931	0.4968	0.3066	0.6141	0.3491	0.8992

B. Photography Images

This section gives the performance of the filters on photography images. Images considered are Lossy, Lossless, and ordinary. Both compressed and uncompressed, colored, and gray scale images. Table VII summarizes Lossless colored photography image results. Negative method produced worst performance for Poisson, Salt and Pepper, Speckle, Average blur, and Gaussian blur distortions measured by SSIM. For Gaussian and low contrast, it is neutral. Digital filter reported bad metrics for PSNR for all kinds of distortions under investigation. Contrast stretching and, Gray-level slicing, and Histogram equalization are neutral for both metrics. De-blur regularized and De-blur blind deconvolution reported bad performance on PSNR for Gaussian noise removal. De-blur blind deconvolution also produced same performance for Speckle noise, while give good performance for Gaussian blur. De-blur Lucy Richardson, Adaptive filtering, and Digital filtering reported different performance that vary from best to neutral performance as seen in the table.

TABLE VII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS COLORED PHOTOGRAPHY IMAGE

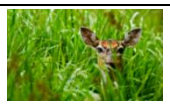
		Enhancement Method									
Distortion	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	4.7224	1.4546	-27.8613	12.327	18.6192	26.6844	9.6147	26.2662	-2.3856	16.6169
	SSIM	-0.5771	0.1072	0.0057	0.566	0.8138	0.9623	0.4299	0.9597	0.196	0.7287
Poisson	***PSNR***	4.6073	12.2438	7.3174	21.7573	24.3296	31,303	10.1437	32.3851	-2.2031	18.7151
	SSIM	-0.6522	0.656	0.0003	0.9347	0.9708	0.9861	0.4424	0.9893	0.2218	0.811
Salt & Pepper	***PSNR***	4.608	4.933	19.5146	16.3255	23.0929	26.2515	10.0394	32.2448	-2.2674	17.9309
	SSIM	-0.6254	0.231	0.8675	0.8023	0.9417	0.9597	0.4362	0.9891	0.2116	0.7787
Speckle	***PSNR***	4.5292	2.0144	-0.4502	12.6783	19.4393	27.2125	9.5283	25.7252	-2.2309	16.8399
	SSIM	-0.589	0.1521	0.2931	0.6515	0.8566	0.968	0.4055	0.9614	0.2044	0.7454
Average Blur	***PSNR***	4.6885	23.1837	7.3174	29.9033	22.3143	27.4588	10.063	28.8997	-2.1112	18.3365
	SSIM	-0.6398	0.9453	0.0003	0.9876	0.9471	0.9701	0.4271	0.9787	0.225	0.7944
Gaussian Blur	***PSNR***	4.6873	23.2716	27.6908	29.9807	22.3564	27.5179	10.065	28.9779	-2.1124	18.3461
	SSIM	-0.6402	0.9461	0.9788	0.9878	0.9478	0.9704	0.4274	0.9791	0.225	0.7948
Low Contrast	***PSNR***	7.6265	13.471	15.1051	15.1045	24.331	14.9448	9.9517	15.093	-3.3598	14.1018
	SSIM	-0.4951	0.5045	0.627	0.6269	0.9708	0.5977	0.4497	0.6135	0.2836	0.5941

TABLE VIII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY COLORED PHOTOGRAPHY IMAGE

		Enhancement Method									
Distortion	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	3.9499	1.3897	-18.696	12.5505	20.0537	26.8877	11.004	26.0529	-1.9452	18.5281
	SSIM	-0.3408	0.0339	0.0494	0.2701	0.571	0.8558	0.1899	0.8222	0.1329	0.5732
Poisson	***PSNR***	3.8931	11.4149	9.2148	21.1978	27.3208	31.125	11.4835	31.6384	-1.7585	22.0905
	SSIM	-0.4682	0.4135	0.4717	0.8266	0.966	0.9451	0.1894	0.9646	0.2308	0.9209
Salt & Pepper	***PSNR***	3.8993	4.6035	-32.5947	16.3535	23.4035	25.4369	11.3163	31.5326	-1.833	20.5724
	SSIM	-0.4247	0.0852	0.7813	0.585	0.8217	0.8301	0.1788	0.9639	0.1875	0.8018
Speckle	***PSNR***	3.8639	2.6723	-43.5257	13.3103	22.0095	26.3354	10.9511	26.066	-1.7724	19.5725
	SSIM	-0.3687	0.0734	0.00040.8792	0.4825	0.8084	0.887	0.1788	0.897	0.1912	0.751
Average Blur	***PSNR***	3.9541	29.5392	24.7359	32.6535	25.2761	27.9628	11.4133	29.3122	-1.69	21.4321
	SSIM	-0.4423	0.9418	0.8792	0.9813	0.9177	0.9141	0.1679	0.944	0.2377	0.8934
Gaussian Blur	***PSNR***	3.9532	29.4786	24.8144	32.6717	25.3149	28.0081	11.4156	29.3777	-1.6908	21.4497
	SSIM	-0.443	0.9411	0.8816	0.9816	0.9192	0.9146	0.1683	0.9447	0.2378	0.8945
Low Contrast	***PSNR***	6.9767	13.0309	14.2979	14.2975	27.3233	14.2067	11.3457	14.2805	-3.2075	15.7015
	SSIM	-0.3077	0.3545	0.5228	0.5227	0.966	0.47	0.1971	0.4907	0.2586	0.6967

Table VIII displays a lossy colored image enhancement with metrics. Negative produced the same performance as with a lossless compressed image. Performance of the filters is almost the same except for De-blur regularized and De-blur blind deconvolution. Where good performance is reported for average blur and Gaussian blur enhancement on PSNR, while none is reported for De-blur blind deconvolution. Fully neutral results are obtained from Gray-level slicing and Histogram equalization.

Results for the uncompressed colored photography image are shown in Table IX. Negative method produced the worst performance for Poisson, Salt and Pepper, Speckle for SSIM as reported for the compressed images. PSNR reported bad Gaussian removal for De-blur regularized and De-blur blind deconvolution. The latter also reported the same performance for Speckle noise removal. While performing well for blurring distortion, as did De-blur Lucy Richardson, contrast stretching works well with Poisson, as did the adaptive filter. For low contrast enhancement, contrast stretching reports good performance on PSNR. Histogram equalization shows the best and good Poisson noise removal. Digital filtering reported bad performance on PSNR for all kinds of distortions. Gray-level slicing gives neutral performance for all distortions.

TABLE IX. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR UNCOMPRESSED COLORED PHOTOGRAPHY IMAGE

		Enhancement Method									
Noise	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	4.6243	-0.138	-26.2324	10.6903	19.5019	23.1905	10.1794	22.247	-2.432	19.6532
	SSIM	-0.4792	0.0407	0.0053	0.3084	0.672	0.8309	0.2862	0.7936	0.1544	0.7078
Poisson	***PSNR***	4.5984	4.4779	17.6254	14.7542	25.0827	25.0994	10.687	23.9455	-2.2659	25.7655
	SSIM	-0.5977	0.2068	0.698	0.6749	0.9156	0.8921	0.3044	0.889	0.2231	0.9574
Salt & Pepper	***PSNR***	4.5998	2.1518	17.3497	13.0764	22.711	23.3416	10.5624	23.8968	-2.3287	22.7091
	SSIM	-0.5584	0.0791	0.6426	0.5293	0.8329	0.8286	0.3	0.8882	0.195	0.8649
Speckle	***PSNR***	4.5456	0.4763	-27.1952	11.3215	20.8075	23.3751	10.1165	22.1733	-2.3193	20.6903
	SSIM	-0.4952	0.0637	0.0875	0.4553	0.7939	0.8624	0.2703	0.8391	0.1895	0.8118
Average Blur	***PSNR***	4.8447	22.4843	25.4931	25.4991	20.9111	21.9051	10.277	22.8752	-2.1029	21.1764
	SSIM	-0.4971	0.8652	0.9428	0.9429	0.7634	0.8123	0.2281	0.8559	0.2319	0.8492
Gaussian Blur	***PSNR***	4.8418	22.4521	25.496	25.499	20.9808	21.9398	10.2876	22.9197	-2.1041	21.248
	SSIM	-0.4996	0.8639	0.943	0.9431	0.7672	0.8134	0.2295	0.8572	0.2323	0.8524
Low Contrast	***PSNR***	7.6169	9.7939	14.4814	14.4689	25.0838	14.6275	20.4852	14.7786	-3.2895	16.3938
	SSIM	-0.3824	0.1999	0.4751	0.4731	0.9156	0,3889	0.2958	0.4202	0.278	0.7285

TABLE X. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS GRAY PHOTOGRAPHY IMAGE

		Enhancement Method									
Noise	**Metric**	***Negative***	***De-blur Regularized Filter***	***De-blur blind deconvolution***	***De-blur Lucy-Richardson***	***Contrast Stretching***	***Adaptive Filtering***	***Gray Level Slicing***	***Median Filtering***	***Digital Filtering***	***Histogram Equalization***
Gaussian	***PSNR***	4.4101	1.5375	6.5716	12.6943	20.7327	29.351	11.1363	27.272	-2.9921	14.3087
	SSIM	-0.0433	0.0053	0.0142	0.0585	0.2171	0.7079	0.1809	0.5282	0.0347	0.1816
Poisson	***PSNR***	4.0705	15.842	25.761	25.7273	37.4614	35.2438	11.618	36.294	-2.9562	15.6659
	SSIM	-0.066	0.3357	0.7975	0.7957	0.9873	0.8764	0.1985	0.9267	0.2973	0.8019
Salt & Pepper	***PSNR***	4.0758	5.1572	16.3423	16.15	24.4369	25.8036	11.4425	36.1679	-3.0009	15.2044
	SSIM	-0.0625	0.0206	0.4074	0.3928	0.7552	0.6907	0.1535	0.926	0.1978	0.6278
Speckle	***PSNR***	4.1571	3.0154	6.5716	13.364	21.7514	26.3883	11.2379	26.688	-2.6785	15.0145
	SSIM	-0.0461	0.0203	0.0142	0.193	0.4464	0.6654	0.1758	0.6267	0.1003	0.3365
Average Blur	***PSNR***	4.0927	22.8789	29.6961	29.6925	32.4192	31.9555	11.5992	13.0391	-2.9284	15.6208
	SSIM	0.0393	0.7712	0.9523	0.9523	0.8003	0.8365	0.1959	0.8902	0.3212	0.7708
Gaussian Blur	***PSNR***	4.0924	22.9726	29.7687	29.7652	32.529	32.0133	11.5998	33.1281	-2.9288	15.6229
	SSIM	0.0371	0.771	0.9529	0.9520	0.9019	0.8367	0.1959	0.8914	0.3219	0.7731
Low Contrast	***PSNR***	7.0918	14.1865	14.6811	14.681	37.4655	14.6572	11.6152	14.6898	-3.1808	15.6412
	SSIM	0.1841	0.439	0.6998	0.6997	0.9873	0.6169	0.2005	0.6628	0.3142	0.8106

In Table X, a lossless gray photography image is tested for insight enhancement performance on the distortions specified. Negative produced the worst SSIM for noise removal. Neutral results for blurring and contrast enhancements. De-blur regularized reported the worst and bad performance for Gaussian noise removal and bad speckle removal. The Digital filter again performs badly based on PSNR. Best results reported for Average blur and Gaussian blur based on PSNR for De-blur blind deconvolution and De-blur Lucy Richardson. Contrast stretching performs well for Poisson removal, blurring, and contrast enhancements. Adaptive filtering for Poisson and Gaussian blurring showed good performance. The median filter also performed well for Poisson, Salt and Pepper, average blur, and Gaussian blur in terms of PSNR. Gray-level slicing and Histogram equalization reported neutral for all distortions.

Table XI reported results for a lossy gray photography image. Negative produced the worst SSIM except for blurring removal. Neutral results reported for Adaptive, Gray-level slicing, and Histogram equalization. Digital filtering reported similar PSNR for distortions. Best and good results scattered for De-blur regularized in blur enhancement for both Average and Gaussian, De-blur Lucy Richardson for blurring enhancement as well, contrast stretching for Gaussian, Poisson, and Salt and Pepper removals, besides low contrast enhancement. Median for Poisson and Salt and Pepper.

TABLE XI. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY GRAY ORDINARY IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	1.7548	-1.7425	3.308	7.4912	21.6734	18.4125	14.7443	19.4072	-7.1631	12. 5738
	SSIM	-0.3663	0.0698	0.161	0.3042	0.7246	0.6305	0.6739	0.7087	0.1585	0.5498
Poisson	***PSNR***	1.515	-0.4764	2.0226	9.5962	12.27244	19.4357	15.1881	20.3165	-7.4613	12.6121
	SSIM	-0.3831	0.1025	0.0227	0.4782	1	0.6925	0.6777	0.8321	0.2499	0.6235
Salt & Pepper	***PSNR***	1.5323	-1.5589	2.1303	9.1207	23.6812	18.9153	14.6518	20.1639	-7.4378	12.2297
	SSIM	-0.3709	0.0781	0.213	0.4311	0.8528	0.6548	0.581	0.8293	0.1905	0.5715
Speckle	***PSNR***	1.9646	-2.5548	-203.6452	8.5801	18.3589	17.25	14.6131	18.1461	-6.7364	12.2699
	SSIM	-0.3671	0.0623	0.0882	0.3779	0.6718	0.603	0.6745	0.6277	0.1706	0.5187
Average Blur	***PSNR***	2.027	25.1248	14.4358	22.5358	18.1736	15.9339	13.3406	17.1045	-7.2574	11.2853
	SSIM	-0.2773	0.9431	0.5199	0.9077	0.7586	0.5428	0.5169	0.6898	0.2834	0.5321
Gaussian Blur	***PSNR***	2.0198	24.7907	13.9909	22.5362	18.3018	15.9774	13.3988	17.1715	-7.259	11.3043
	SSIM	-0.2807	0.9315	0.5171	0.9082	0.7661	0.5453	0.5209	0.6938	0.2846	0.5373
Low Contrast	***PSNR***	4.613	6.4466	11.6897	11.6599	24.6906	11.2021	15.1881	11.6943	-4.283	11.9894
	SSIM	-0.3069	0.1964	0.6062	0.6017	1	0.4093	0.6777	0.5138	0.4373	0.6908

TABLE XII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR AN UNCOMPRESSED GRAY ORDINARY IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	3.5858	-0.8225	8.0105	10.6085	20.8104	21.7001	11.132	21.0616	-1.6251	12.3396
	SSIM	-0.3686	0.0308	0.0553	0.2512	0.6273	0.574	0.2638	0.6112	0.127	0.4773
Poisson	***PSNR***	3.4763	2.1035	16.4471	13.3667	32.9088	22.9341	11.6328	22.3551	-1.4307	13.162
	SSIM	-0.4784	0.0726	0.4754	0.5259	0.9971	0.6733	0.2687	0.7522	0.2406	0.6513
Salt & Pepper	***PSNR***	3.4839	0.4871	-25.99815	12.2545	24.2201	21.8025	11.4502	22.3003	-1.5108	12.9193
	SSIM	-0.4467	0.047	0.0412	0.4505	0.8915	0.618	0.2525	0.7505	0.2047	0.5988
Speckle	***PSNR***	3.4501	-0.2055	7.4218	11.2676	22.4115	22.0827	11.0975	21.1847	-1.4634	12.8447
	SSIM	-0.3878	0.0423	0.2674	0.4324	0.8568	0.6655	0.2518	0.7021	0.205	0.5548
Average Blur	***PSNR***	3.7517	25.1915	24.9368	24.9475	20.5853	20.3003	11.0766	21.2661	-1.1765	12.639
	SSIM	-0.2897	0.8465	0.89	0.8904	0.7398	0.5539	0.1754	0.6748	0.2648	0.5081
Gaussian Blur	***PSNR***	3.7484	25.0361	24.9733	24.975	20.6828	20.3288	11.0893	21.3113	-1.1782	12.658
	SSIM	-0.2944	0.842	0.8915	0.8918	0.7475	0.5555	0.1768	0.6781	0.2661	0.5134
Low Contrast	***PSNR***	6.5629	8.3607	13.5009	13.4769	32.9143	13.4127	11.5911	13.5538	-3.0227	13.1525
	SSIM	-0.3521		0.4551	0.4498	0.9971	0.2757	0.2829	0.3326	0.2315	0.6535

Table XII displays an uncompressed gray ordinary image result. Negative performs worst except for blurring. Digital reported bad results for distortions. Neutral results reported for Gray-level, Median, and Histogram equalization. For blurring enhancement, the performance of De-blur regularized, De-blur blind deconvolution, and De-blur Lucy Richardson was reported to be good on PSNR and best on SSIM for the last two. Contrast stretching performs well for Poisson and Salt and Pepper noise removal, in addition to low contrast enhancement.

C. Macro Images

Section C reported results of the 10 filters on lossless, lossy, and uncompressed colored satellite images. The compressed and uncompressed gray satellite images are considered as well.

Table XIII reported lossless colored macro image enhancement results. Best blurring enhancement achieved by De-blur regularized, De-blur blind deconvolution, and De-blur Lucy Richardson. Median produced a good Gaussian blurring enhancement. Contrast enhancement best results are reported for contrast stretching. Digital filtering performs badly based on PSNR. Negative reported the worst results for Poisson, Salt and Pepper, Average blur, Gaussian blur, and low contrast.

TABLE XIII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS COLORED MACRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	8.5245	0.2456	7.1741	10.6863	17.2548	23.8054	8.1926	23.7806	-2.5233	15.9065
	SSIM	-0.4261	0.0259	0.0002	0.2061	0.5299	0.6831	0.1771	0.6989	0.1204	0.4975
Poisson	***PSNR***	8.8066	7.6821	17.6305	17.4386	21.5765	25.5937	8.7588	26.6837	-2.3311	16.4511
	SSIM	-0.6692	0.1266	0.5908	0.5791	0.9398	0.7532	0.2032	0.849	0.2231	0.7586
Salt & Pepper	***PSNR***	8.756	4.1033	14.5551	14.3149	21.2891	24.1781	8.7114	26.6162	-2.3961	16.3787
	SSIM	-0.5866	0.0613	0.4101	0.3944	0.8328	0.7026	0.1996	0.8475	0.1824	0.6928
Speckle	***PSNR***	8.5717	1.1222	7.1741	11.8711	19.308	23.956	8.2252	23.6582	-2.4725	16.6152
	SSIM	-0.4619	0.0319	0.0002	0.2714	0.6004	0.6959	0.1649	0.7032	0.1368	0.5424
Average Blur	***PSNR***	9.2598	25.8571	29.3332	29.3331	19.9576	23.3933	8.5885	25.0953	-2.2047	15.7452
	SSIM	-0.4862	0.9019	0.9365	0.9365	0.8386	0.6587	0.1528	0.7843	0.2522	0.6999
Gaussian Blur	***PSNR***	9.2536	25.8838	29.3725	29.37	19.9952	23.424	8.5939	25.1502	-2.205	15.7611
	SSIM	-0.4906	0.9005	0.9371	0.9371	0.8429	0.6602	0.1539	0.7868	0.2525	0.7036
Low Contrast	***PSNR***	11.5668	12.0807	18.3731	18.3591	21.5779	18.3349	8.4636	18.7731	-3.8061	15.5871
	SSIM	-0.5134		0.675	0.6731	0.9398	0.5401	0.2479	0.6473	0.2892	0.7387

TABLE XIV. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY COLORED MACRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	9.4141	0.7991	11.3507	11.4129	16.9719	27.7898	8.1516	26.3343	-2.3683	14.7554
	SSIM	-0.2513	0.0119	0.1089	0.1097	0.2797	0.7823	0.0868	0.6974	0.0811	0.2802
Poisson	***PSNR***	9.8063	13.7843	23.9378	23.8913	17.3057	31.9443	8.6961	32.9808	-2.1631	14.8163
	SSIM	-0.5489	0.2333	0.7597	0.758	0.6896	0.8943	0.1108	0.9358	0.2273	0.6374
Salt & Pepper	***PSNR***	9.7356	5.853	16.2159	16.1125	16.8585	26.7697	8.6509	32.9067	-2.2331	14.8035
	SSIM	-0.4394	0.0399	0.3394	0.3331	0.5708	0.7478	0.1065	0.9349	0.1665	0.5364
Speckle	***PSNR***	9.519	2.0696	12.5451	12.7044	16.2623	27.3054	8.2375	26.0518	-2.3068	15.3003
	SSIM	-0.2897	0.0157	0.148	0.152	0.3296	0.8015	0.0782	0.7189	0.1049	0.3155
Average Blur	***PSNR***	9.9547	26.508	33.0022	33.0003	17.4427	29.1415	8.6358	30.7264	-2.1159	14.6159
	SSIM	-0.4497	0.8831	0.9643	0.9643	0.6829	0.8405	0.092	0.8983	0.2427	0.6454
Gaussian Blur	***PSNR***	9.9527	26.6172	33.1031	33.1007	17.4496	29.1857	8.6376	30.7949	-2.1165	14.6194
	SSIM	-0.452	0.8836	0.9647	0.9647	0.6846	0.8415	0.0924	0.8997	0.2428	0.6465
Low Contrast	***PSNR***	11.116	15.9929	23.68	23.6689	17.305	24.7439	8.6653	25.2476	-2.5862	14.2612
	SSIM	-0.3071	0.2351	0.6714	0.6707	0.6895	0.6467	0.108	0.7101	0.2675	0.5831

Table XIV reported lossy colored macro image results. Best luring enhancement reported for De-blur blind deconvolution, and De-blur Lucy Richardson. Adaptive and Median filtering also reported good performance on PSNR. Noise removal achieved the best SSIM for Median on Poisson noise, while PSNR reported good results for passion, Salt and Pepper, Average blur, and Gaussian blur. Digital filter reported weak PSNR for all distortions. De-regularized reports similar performance for Gaussian, Salt and Pepper, and Speckle noise removal.

Uncompressed colored macro image results are shown in Table XV. Negative filter performance was reported as worst on Poisson, Salt and Pepper, Average, Gaussian blur, and Low contrast. The Digital filter also performs badly except for the Average blur. For blurring enhancement, De-blur Richardson reported good and best performance on PSNR and SSIM, respectively. Contrast stretching achieved the best noise removal based on SSIM for Poisson noise, and the best SSIM for low contrast enhancement. Adaptive filter performs well for Gaussian, Poisson, Salt and Pepper, and Speckle based on PSNR.

TABLE XV. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR UNCOMPRESSED COLORED MACRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	6.9584	0.3041	21.3381	11.1216	17.7685	25.7688	8.5676	23.9856	-2.6037	16.9936
	SSIM	-0.3629	0.0176	0.0112	0.1606	0.4456	0.6968	0.2177	0.6661	0.1026	0.4652
Poisson	***PSNR***	6.9929	6.8399	22.1874	17.4137	18.6496	28.4654	9.1077	26.4352	-2.4183	18.2794
	SSIM	-0.6218	0.1257	0.808	0.6284	0.9399	0.8286	0.243	0.8536	0.2571	0.7915
Salt & Pepper	***PSNR***	6.9712	3.556	-28.7838	14.3469	19.3795	25.2719	9.045	26.4031	-2.4793	17.8612
	SSIM	-0.5322	0.0431	0.015	0.3917	0.7919	0.7133	0.2	0.8525	0.1818	0.6852
Speckle	***PSNR***	6.8372	0.9052	-3.1781	11.7186	19.6912	25.6903	8.568	23.6576	-2.5583	17.6717
	SSIM	-0.3916	0.0264	0.4299	0.2879	0.5921	0.7646	0.2058	0.7154	0.1709	0.5045
Average Blur	***PSNR***	7.1862	24.7441	-14.959	27.5731	16.6835	25.1337	8.9266	25.8319	-2.3567	17.2665
	SSIM	-0.4473	0.8332	0.7975	0.9289	0.7994	0.7523	0.1992	0.8119	0.27	0.7201
Gaussian Blur	***PSNR***	7.184	24.6481	7.0942	27.577	16.7035	25.1795	8.9328	25.875	-2.357	17.2876
	SSIM	-0.4513	0.8302	0.0774	0.9295	0.8041	0.7534	0.2001	0.8136	0.2706	0.7245
Low Contrast	***PSNR***	10.04	11.589	16.1181	16.1134	18.6509	16.5368	8.3562	16.5271	-4.1663	16.3612
	SSIM	-0.4234	0.1702	0.5567	0.5554	0.9399	0.4987	0.2847	0.5445	0.2769	0.7097

TABLE XVI. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSY GRAY MACRO IMAGE

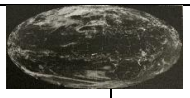


Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	6.2452	-3.3558	12.7743	8.8843	19.3859	19.2621	9.3728	18.1566	-1.6874	13.0794
	SSIM	-0.5512	0.0282	0.4866	0.2578	0.6997	0.4472	0.2897	0.4358	0.1431	0.5685
Poisson	***PSNR***	6.3657	-1.6396	14.8625	11.0544	30.6998	20.1136	9.8726	18.7738	-1.4563	14.3647
	SSIM	-0.6917	0.0419	0.5727	0.4201	0.9881	0.4964	0.3087	0.5187	0.2212	0.8139
Salt & Pepper	***PSNR***	6.3512	-2.2411	13.9067	10.4148	24.9784	19.6691	9.7866	18.7393	-1.5477	14.1549
	SSIM	-0.6517	0.0364	0.2897	0.3669	0.9192	0.4724	0.3058	0.5171	0.1969	0.7738
Speckle	***PSNR***	6.2812	-2.781	12.3243	9.8708	21.3404	19.5612	9.4096	18.2706	0.1706	13.599
	SSIM	-0.5868	0.032	0.52	0.3247	0.8143	0.4814	0.2779	0.4737	0.1706	0.6596
Average Blur	***PSNR***	7.3757	22.4106	21.4357	21.4453	15.3818	17.7486	8.8322	18.2377	-1.1086	12.4574
	SSIM	-0.2459	0.8332	0.7747	0.7757	0.5669	0.3111	0.1098	0.42	0.2432	0.5348
Gaussian Blur	***PSNR***	7.365	22.2428	21.4991	21.5013	15.487	17.7735	8.863	18.2658	-1.1089	12.5124
	SSIM	-0.2562	0.827	0.7794	0.7802	0.5816	0.3134	0.114	0.4239	0.2471	0.5485
Low Contrast	***PSNR***	9.2984	5.0334	15.2332	14.9748	30.7015	15.4942	9.8266	15.4786	-3.1398	14.3443
	SSIM	-0.5106	0.0839	0.5753	0.5425	0.9881	0.2809	0.3291	0.3116	0.2774	0.8129

Table XVI shows results for the lossy gray macro image. Negative method produced the worst performance for Gaussian, Poisson, Salt and Pepper, Speckle, and low contrast based on SSIM. The Digital filter also performs badly except for the Average blur. De-blur regularized works well for blurring enhancement. Contrast stretching performs well for Poisson, Salt and Pepper, Speckle, and Low contrast enhancements. Neutral results are obtained for all distortions reported for Adaptive, Median, Gray-level Slicing, and Histogram equalization.

Table XVII reported enhancement results for a lossless gray macro image. Negative filter reported the worst SSIM for noise removal on all noise types, in addition to Gaussian blur enhancement. Performance of the Digital filter PSNR reported bad performance against others. Besides De-blur regularized performance for Gaussian and Speckle noise removal, and De-blur blind deconvolution for Salt and Pepper noise removal. For blurring enhancement, De-blur Lucy Richardson reported the best and good SSIM and PSNR, respectively. Contrast stretching performs well for Poisson noise and Low contrast enhancements. Adaptive filtering did well for all noise kinds. Median filtering competes for Poisson and Salt and Pepper noise removal.

TABLE XVII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR LOSSLESS GRAY MACRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	3.2583	0.3299	9.5034	12.0956	20.2881	24.1073	11.8037	22.8442	0.235	10.3059
	SSIM	-0.2632	0.0285	0.1753	0.2048	0.4577	0.5948	0.3464	0.6007	0.0958	0.3097
Poisson	***PSNR***	3.1287	4.8491	9.5034	18.3003	29.7177	26.9783	12.2716	24.493	0.0518	10.7031
	SSIM	-0.3652	0.0886	0.1753	0.7574	0.9928	0.8161	0.353	0.865	0.3628	0.5255
Salt & Pepper	***PSNR***	3.138	2.1709	26.582	15.6214	23.9044	24.4145	12.0556	24.4456	0.0764	10.6011
	SSIM	-0.3361	0.0508	0.1747	0.5692	0.8502	0.7254	0.3041	0.8637	0.2716	0.4773
Speckle	***PSNR***	3.1064	1.9612	12.4028	14.6432	23.8268	25.0197	11.7976	23.2871	0.0034	10.557
	SSIM	-0.286	0.0423	0.3601	0.5405	0.8206	0.7907	0.3365	0.7966	0.3088	0.4047
Average Blur	***PSNR***	3.2936	23.6974	9.5034	28.957	22.1945	22.5311	11.8549	23.4673	0.3628	110.5401
	SSIM	-0.2565	0.7706	0.1753	0.9511	0.8583	0.7085	0.2783	0.8098	0.3843	0.448
Gaussian Blur	***PSNR***	3.2914	23.3367	23.0032	29.1982	22.2817	22.5839	11.8641	23.5119	0.3598	10.547
	SSIM	-0.2595	0.76	0.8007	0.9554	0.8629	0.7103	0.2791	0.8121	0.3848	0.4508
Low Contrast	***PSNR***	6.447	9.8478	13.3459	13.3381	29.7213	13.2368	12.136	13.3517	2.8746	11.0058
	SSIM	-0.2422	0.1458	0.4862	0.4842	0.9928	0.3494	0.3765	0.3952	0.204	0.5282

TABLE XVIII. PERFORMANCE OF ENHANCEMENT METHODS ON DISTORTION TYPES FOR UNCOMPRESSED GRAY MACRO IMAGE



Noise	Metric	Enhancement Method									
		Negative	*De-blur Regularized Filter*	*De-blur blind deconvolution*	*De-blur Lucy-Richardson*	*Contrast Stretching*	*Adaptive Filtering*	*Gray Level Slicing*	*Median Filtering*	*Digital Filtering*	*Histogram Equalization*
Gaussian	***PSNR***	4.7725	-0.6596	12.8548	9.3736	20.4373	18.6575	10.5562	20.5453	-4.2811	19.327
	SSIM	-0.5875	0.0901	0.4122	0.4301	0.8115	0.6591	0.4814	0.8045	0.1792	0.793
Poisson	***PSNR***	4.7934	3.2133	16.3253	12.8013	12.33671	19.3428	11.0309	21.9149	-4.2384	25.5406
	SSIM	-0.684	0.202	0.7562	0.6638	1	0.6926	0.5062	0.8665	0.2219	0.9652
Salt & Pepper	***PSNR***	4.7928	1.3128	-29.6138	11.3039	24.714	18.9848	10.8913	21.8395	-4.2612	22.1598
	SSIM	-0.6535	0.1421	0.6136	0.5819	0.9421	0.6761	0.4951	0.8647	0.2085	0.9152
Speckle	***PSNR***	4.8409	-0.698	12.4154	9.9906	20.149	18.5017	10.498	19.9541	-4.1042	19.5778
	SSIM	-0.5972	0.0883	0.4919	0.4956	0.839	0.6594	0.4676	0.7935	0.1921	0.8199
Average Blur	***PSNR***	5.493	30.0563	27.3753	27.3827	21.316	17.4243	10.5677	19.5291	-3.9327	20.1496
	SSIM	-0.5669	0.9748	0.9661	0.9662	0.8557	0.5747	0.4266	0.7601	0.2611	0.8467
Gaussian Blur	***PSNR***	5.4819	29.5567	27.4728	27.4607	21.4674	17.4512	10.5893	19.5875	-3.936	20.2649
	SSIM	-0.5709	0.9715	0.9672	0.9671	0.861	0.5766	0.429	0.7634	0.2611	0.8515
Low Contrast	***PSNR***	7.8914	10.2821	15.606	15.5919	24.4513	13.8039	11.0309	14.5404	-3.4682	25.4996
	SSIM	-0.5056	0.3443	0.7261	0.7249	1	0.3535	0.5062	0.4987	0.3626	0.9645

Table XVIII displays enhancement methods on an uncompressed gray macro image. Negative filter reported the worst SSIM for noise removal on all noise types, in addition to Gaussian blur enhancement. The performance of the Digital filter PSNR was reported as bad. Best blurring enhancement reported for De-blur regularized, De-blur blind deconvolution. For contrast enhancement, contrast stretching and histogram equalization perform superior on others.

The tables for the three categories are further visualized in bar graphs that demonstrate the effects of the 10 filters on the 7 distortions.

Fig. 5 visualizes the PSNR that measures the effectiveness of the filters on the distortions over the micro images. The category includes Lossy, Lossless, both colored and uncolored, and compressed and uncompressed. The filters shown in the legend are indicated by the colors that distinguish each.

The PSNR measures the effectiveness of the filters on the distortions over the ordinary images (Fig. 6). The category includes the Lossy, Lossless, both colored and uncolored, and compressed and uncompressed. The filters shown in the legend are indicated by the colors that distinguish each.

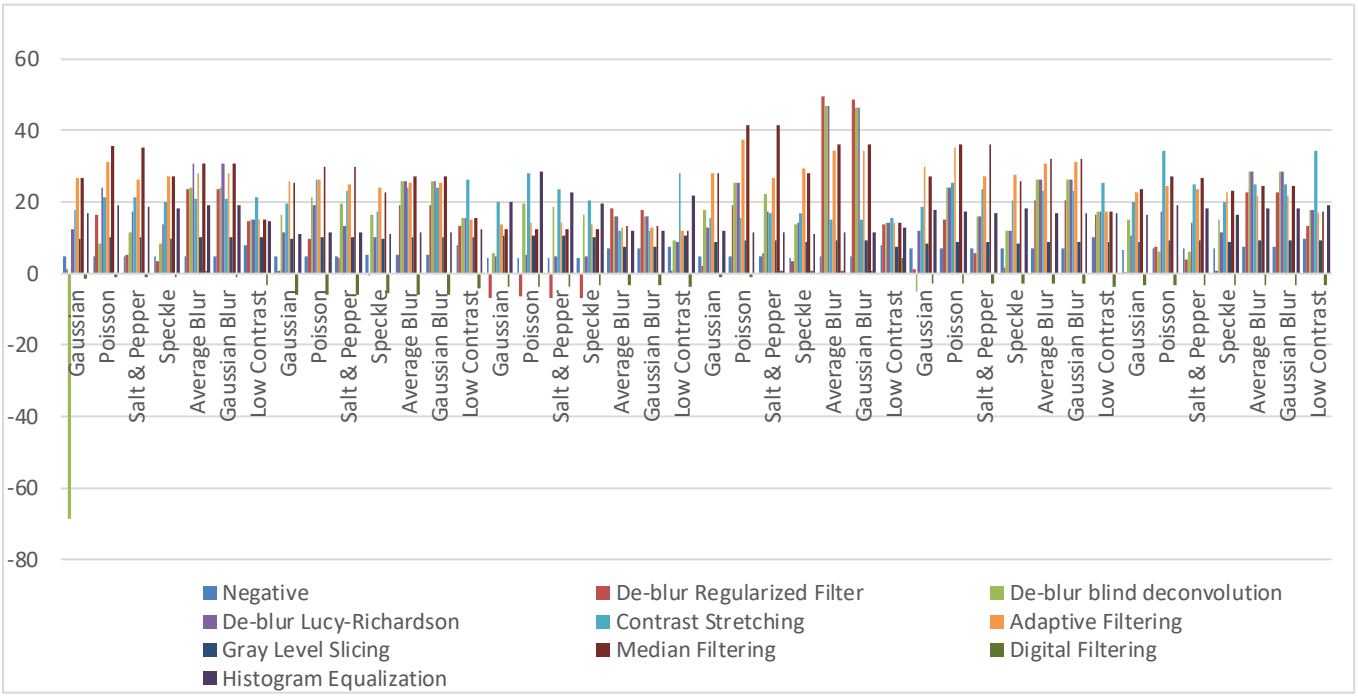


Fig. 5. Performance of enhancement filters on the distortion for micro images.

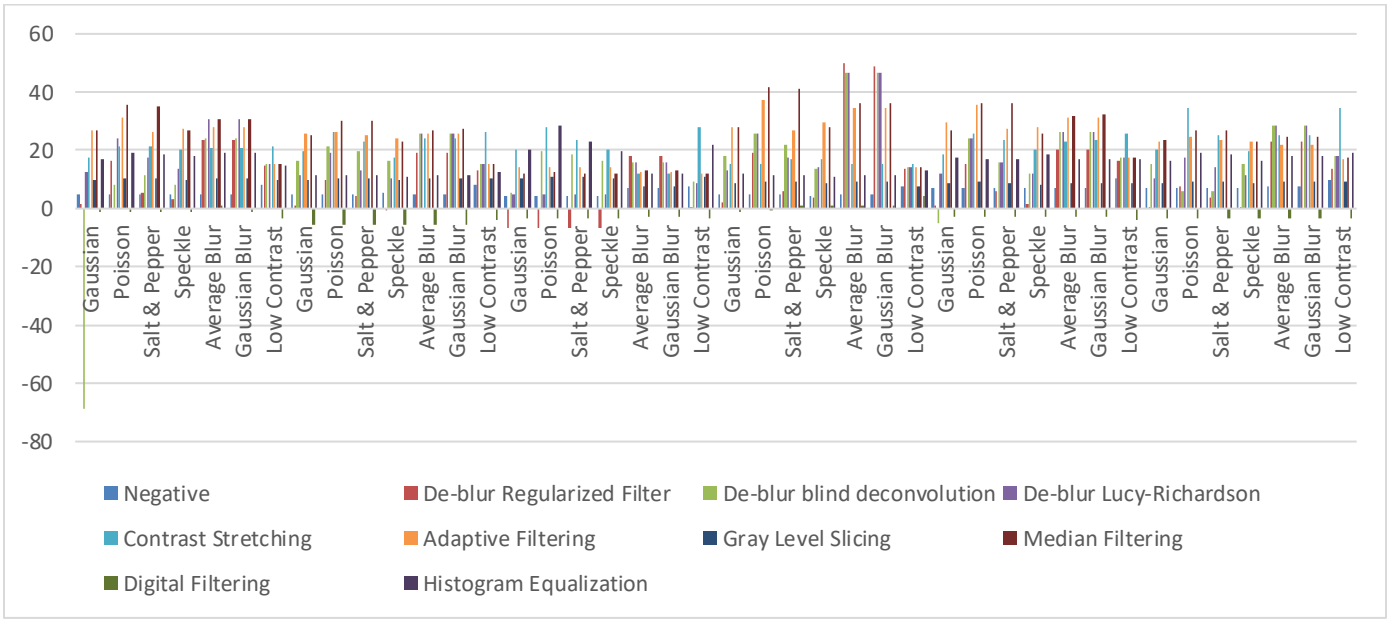


Fig. 6. Performance of enhancement filters on the distortion for ordinary images.

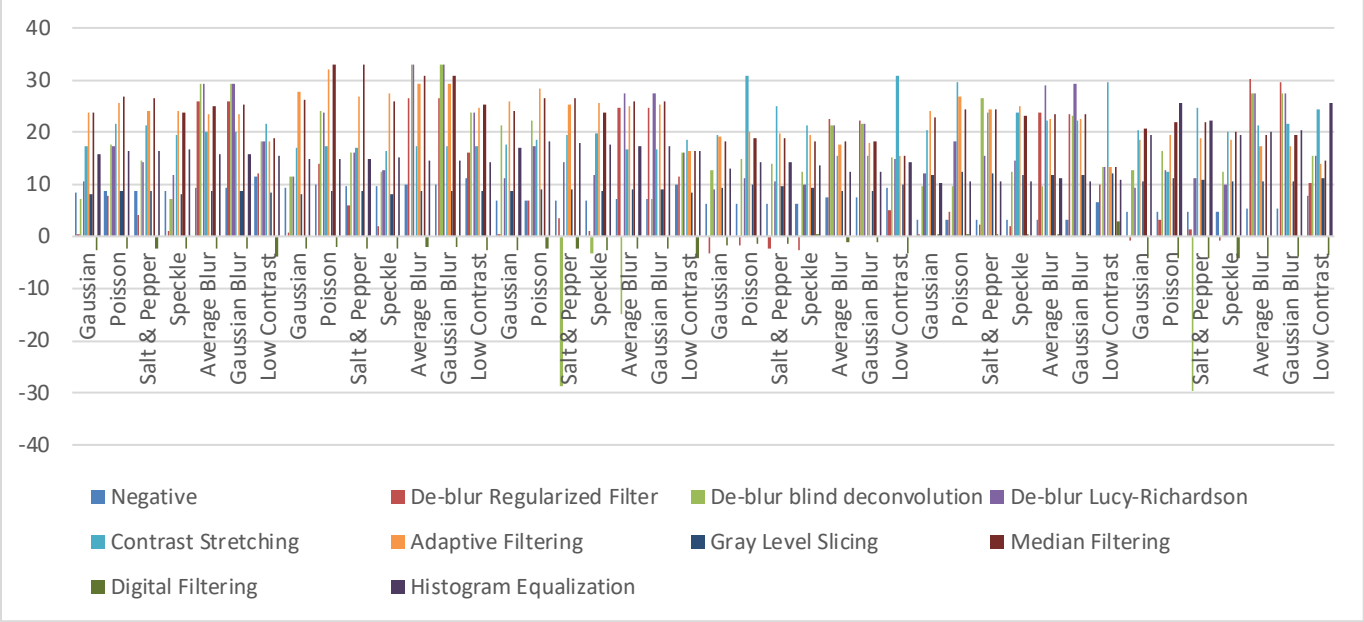


Fig. 7. Performance of enhancement filters on the distortion for macro images.

Fig. 7 visualizes the PSNR that measures the effectiveness of the filters on the distortions over the macro images. The category includes Lossy, Lossless, both colored and uncolored, and compressed and uncompressed. The filters shown in the legend are indicated by the colors that distinguish each.

VI. Conclusions and Recommendations

In this study, an investigation and comparison of Negative, De-blur Regularized, De-blur blind deconvolution, De-blur Lucy-Richardson, Contrast Stretching, Adaptive, Gray Level Slicing, Median, Digital, and Histogram Equalization filters on distortions are applied to micro, macro, and photography images. Key performance indicators that judge the performance are Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Section V gives the findings and results for the 10 enhancement methods. Negative filter fails to enhance the micro, photography, and macro images based on the SSIM metric, while producing neutral results for PSNR. De-blur Regularized filter sometimes reported good performance for blurring enhancement, but overall, across the three categories, it fluctuates and tends towards the worst category (Table XIX). Digital filtering reported bad PSNR for all three categories. De-blur blind deconvolution failed to compete for higher achievement with other filters. Besides gray-level slicing and Histogram equalization, these methods may be more effective in specific applications. The best methods reported for all three categories, whether compressed or uncompressed, colored, or gray, are reflected in Table XIX. It is found that De-blur, Richardson, Median, and Contrast stretching are the best among others. To strengthen the performance, hybrid filtering may introduce more enhancement capabilities.

De-blur Lucy-Richardson, Contrast Stretching, and Media Filtering have the best performance over all image types. Negative and Digital filtering have the worst performance over all image types. Histogram Equalization and Grey Level Slicing always give a moderate performance in all image types, but it has a good result if the image is not compressed. De-blur Regularized Filter and De-blur blind deconvolution have good performance when the image suffers from Blur distortion. Contrast Stretching has good performance when the image suffers from low contrast.

TABLE XIX. Recommended Enhancement Methods for Micro Versus Macro and Ordinary Images

Image Type	Best Enhancement Method	Worst Enhancement Method
Mico (Medical)	De-blur Lucy-Richardson	Negative
	Median Filtering	Digital Filtering
	Contrast Stretching	De-blur Regularized Filter
Ordinary Photography	De-blur Lucy-Richardson	Negative
	Contrast Stretching	Digital Filtering
	Median Filtering	De-blur Regularized Filter
Macro (Satellite)	De-blur Lucy-Richardson	Negative
	Contrast Stretching	Digital Filtering
	Median Filtering	De-blur Regularized Filter

Acknowledgment

The authors would like to acknowledge Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R925), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] Fang Xu, Hui Zhang, Xue-Rui Ai, Jin Cao, Xin-Yu Xu, "Research on image enhancement preprocessing techniques," Proc. SPIE 13690, International Conference on Advances in Computer Vision Research and Applications (ACVRA 2025), 1369002 (4 July 2025) https://doi.org/10.1117/12.3068659.

[2] Jiawei Guo, Jieming Ma, Ángel F. García-Fernández, Yungang Zhang, Haining Liang "A survey on image enhancement for Low-light images," Heliyon 9 (2023) e14558; Elsevier Ltd. https://doi.org/10.1016/j.heliyon.2023.e14558.

[3] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, Kanta Prasad Sharma, Deena Nath Gupta, "A deep journey into image enhancement: A survey of current and emerging trends", Information Fusion, Volume 93, 2023, Pages 36-76, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.12.012.

[4] Liu Z. et al. "Adaptive Super-Resolution Networks for Single Imaging at Ultra-Low Sampling Rate" 2024, IEEE Access, Vol 12, pp 78496-78504, doi:10.1109/ACCESS 2024.34-2693.

[5] Zhen Tian, Peixin Qu, Jielin Li, Yukun Sun, Guohou Li, Zheng Liang and Weidong Zhang " A Survey of Deep Learning-Based Low-Light Image Enhancement". Sensors 2023, 23, 7763. https://doi.org/10.3390/s23187763.

[6] Jonathan Sanderson, Syed Rafay Hasan, "A Case Study of Image Enhancement Algorithms Effectiveness of Improving Neural Networks Performance on Adverse Images", arXiv:2312.09509 v2 [eess.IV] 18 Dec 2023.

[7] Hewa Majeed Zangana1, Firas Mahmood Mustafa, Ayaz Khalid Mohammed, Naaman Omar, "Enhancing Image Quality with Deep Learning: Techniques and Applications", Vol. 8, No. 2, December 2024, page. 119-131, ISSN 2598-3245 (Print), ISSN 2598-3288 (Online); DOI: http://doi.org/10.31961/eltikom.v8i2.1242.

[8] Peipei Zhang, "Image Enhancement Method Based on Deep Learning", Hindawi Mathematical Problems in Engineering, Volume 2022, Article ID 6797367, 9 pages; https://doi.org/10.1155/2022/6797367.

[9] Yunliang Qi, Zhen Yang , Wenhao Sun, Meng Lou, Jing Lian, Wenwei Zhao, Xiangyu Deng, Yide Ma, " A Comprehensive Overview of Image Enhancement Techniques", Archives of Computational Methods in Engineering, © CIMNE, Barcelona, Spain 2021 https://doi.org/10.1007/s11831-021-09587-6.

[10] Patel Omprakash, Maravi Yogendra and Sharma Sanjeev "A Comparative Study of Histogram Equalization Based Image Enhancement Techniques for Brighter Preservation and Contrast Enhancement", Signal and Image Processing. International Journal (SIPU)Vol 4, No. 5, October 2013

[11] Khadise Swati and Nagori Meghana, " Implementation and Comparison of Image Enhancement Techniques", International Journal of Computer Applications (0975-8887), Vol. 96-No. 4, June 2014.

[12] C. Dong, C. Loy, K. He and X. Tang." Image Super Resolution Using Deep Convolusion Neural Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, no 2, pp 295-307 2016, DOI 10.1109/TPAMI 2015.2439281.

[13] C. Ledig et al. "Phot-Realistic Single Image Super-Resolution Using a Generative Adversarial Network" 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Honolulu, HI, USA 2017, pp 105-114, doi: 10.1109/CVPR 2017.19.

[14] Wang Z., Bovik A. C., Sheikh H. R., and Simoncelli E. P. " Image Quality Assessment from Error Visibilty to Structural Similarity" IEEE Transactions on Image Processing, 2004.

[15] Gonzalez Rafael and Woods Richard, " Digital Image Processing", 2018, Pearson, ISBN 10: 1-292-22304-9, ISBN 13: 978-1-292-22304-9.

[16] Jian A. K. "Fundamentals of Digital Image Processing", 1989, Prentice Hall.

[17] Pitas I, "Digital Image Processing Algorithms and Applications ", Wiley, 2000.

[18] Szeliski Richard, " Computer Vision Algorithms and Applications" 2022, Springer, ISBN: 978-3-030-34371-2, ISBN 978-3-030-3472-9. doi 10.1007/978-3-030-34372-9.

A Hybrid Neuro-Fuzzy and Machine Learning Model for Student Performance Prediction in Online Mathematics Education

Beyza Esin Özseven[1], Turgut Özseven[2]
Republic of Turkey, Ministry of National Education, Tokat, Turkiye[1]
Department of Computer Engineering, Tokat Gaziosmanpasa University, Tokat, Turkiye[2]

***Abstract*—Accurate and early prediction of student success in online mathematics education is critical for improving learning processes and developing personalized instruction strategies. However, students' problem-solving behaviors, interaction levels, and learning speeds in online learning environments inherently contain uncertainty, limiting the effectiveness of traditional assessment and singular machine learning approaches. This study proposes a hybrid neuro-fuzzy and machine learning-based student success prediction framework that combines the ability of fuzzy logic to represent uncertainty with the strong generalization and adaptive learning capabilities of machine learning methods. In the proposed approach, student success trends are primarily modeled using ANFIS, Random Forest, and XGBoost models, employing raw and derived features on the ASSISTments dataset. The predictions from these models are treated as continuous success representations reflecting uncertainty in students' learning behaviors and are used as input to a hybrid classification structure to make binary success/failure decisions. Thus, the ANFIS model is positioned as an uncertainty-aware and interpretable context generator, while the Random Forest and XGBoost models provide discriminative classification power. Experimental results demonstrate that both ANFIS and Random Forest models exhibit high individual performance; however, combining their complementary features within a hybrid structure significantly increases prediction stability and generalization. Unlike the limitations of 'black box' models in the literature, the interpretability provided by ANFIS, through its linguistic rules and membership functions, enables the proposed approach to generate pedagogically transparent and actionable implications. The findings reveal that this hybrid framework, which integrates uncertainty management with high prediction accuracy, offers a powerful decision-support mechanism for early identification of at-risk students and for developing personalized intervention strategies in online mathematics education.**

***Keywords*—ANFIS; ensemble learning; fuzzy logic; hybrid models; learning analytics**

I. Introduction

The development of online learning platforms has revolutionized educational analytics, enabling the massive collection and analysis of student behavior [1], [2], [3], [4]. These platforms have become digital ecosystems that enable monitoring students' interactions with the learning process, not only in terms of outcomes but also in terms of the process itself. These digital ecosystems record each student's interactions with the system in the form of time-stamped behavioral logs, such as clickstreams, resource viewing times, and forum participation [1, 5, 6]. This type of data reveals not only what students have learned, but also how and under what conditions they have learned. In particular, intelligent tutoring systems such as ASSISTments [7] reflect students' cognitive states during mathematical problem-solving through multidimensional attributes, such as correct/incorrect answer rates, hint requests, and the number of attempts required to solve the problem [1, 6, 8]. In this respect, the ASSISTments dataset provides a rich, high-resolution dataset for predicting student achievement at an early stage.

Mathematics education, by its very nature, exhibits a high degree of logical interdependence and a cumulative structure. A failure to fully understand a fundamental concept increases the student's risk of subsequent failure [9]. Therefore, predicting student achievement early, before end-of-term assessments are completed, is a critical requirement for educators to intervene promptly.

The literature shows that early warning systems significantly increase student retention and overall success through personalized learning paths, targeted feedback, and supportive content [1, 5, 10 - 14]. However, the effectiveness of these systems depends on the extent to which the forecasting models used can represent uncertainty and individual differences.

Online student behavior exhibits ambiguity, variability, and fuzzy patterns stemming from the nature of human reasoning [15 - 17]. Students' moderate level of engagement or cognitive stability in the problem-solving process cannot be adequately represented by traditional models built on sharp boundaries and linear assumptions. This leads to limitations not only in classical statistical methods but also in machine learning models that prioritize accuracy.

Although artificial neural networks and tree-based learning algorithms are successful at capturing complex patterns, many of these models are criticized for being "black boxes" in terms of the interpretability of decision-making processes [18 - 19]. In the field of education, producing pedagogically meaningful and explainable outcomes is a fundamental requirement, alongside the model's accuracy.

In this context, hybrid neurofuzzy approaches, which combine the uncertainty modeling capabilities of fuzzy logic with the data-learning capacity of machine learning, are among the most promising solutions for predicting student achievement [11, 20]. In particular, the Adaptive-Network-based Fuzzy

Inference System (ANFIS) offers educators a significant advantage by providing high predictive performance through an interpretable structure based on IF-THEN rules [4, 11, 21, 22].

This study proposes a hybrid neurofuzzy and machine learning-based model for predicting student success in online mathematics education using the ASSISTments dataset. Extensive experiments were conducted using ANFIS, Random Forest (RF), and XGBoost models with raw and derived student features. The resulting continuous success representations were integrated into a binary classification structure to determine the risk of student failure. The main contribution of this study is to effectively model uncertainty in online mathematics education, to provide a balanced trade-off between interpretability and predictive performance, and to demonstrate the applicability of the proposed hybrid approach for early warning systems.

The remaining sections of the study are structured as follows: The second section summarizes the relevant literature and research gaps. The third section details the ASSISTments dataset, data preprocessing, and feature engineering. The fourth section presents the proposed hybrid ANFIS-RF-XGBoost architecture and learning processes. The fifth section presents the experimental findings with a comprehensive analysis, and the sixth section discusses these findings in light of the literature, outlining the study's limitations and suggesting future research.

II. Related Works

Predicting student academic performance has witnessed a paradigm shift in the fields of educational data mining and learning analytics over the last decade, moving from statistical models to hybrid intelligent systems [23]. Studies in the literature can be categorized primarily by model architecture, dataset diversity, and interpretability of results.

A. Traditional Machine Learning and Ensemble Models

Early educational data mining studies generally focused on singular algorithms such as decision trees, support vector machines, and logistic regression [1, 3]. However, recent research has shown that ensemble learning methods provide higher accuracy than single classifiers [5, 24]. For example, the success of the XGBoost algorithm in capturing complex, nonlinear relationships in structured training data has been highlighted in numerous studies [4, 25]. Similarly, the RF model has been reported to exhibit robust performance, particularly in addressing feature selection and inter-class imbalance issues [26 - 28]. Ponnan et al. (2025) achieved a test accuracy of 98.06% in student assessment using AdaBoost and RF models [11].

B. ANFIS Models

ANFIS models are gaining prominence for modeling uncertainties in educational processes and for handling fuzzy data based on human reasoning [22]. ANFIS is described as a universal estimator that combines the learning ability of artificial neural networks with the uncertainty management capacity of fuzzy logic [6, 29 - 33]. Taylan and Karagozoglu (2009) stated that the ANFIS model yields results as robust as statistical methods and offers a more natural language for interpreting student outcomes [34]. Hybrid models developed in recent years, such as FCM-PSO-ANFIS, have reduced error rates by fine-tuning parameters using metaheuristic optimization algorithms [35].

C. Deep Learning and Time Series Analysis

The timestamped nature of student interaction data has encouraged the use of RNN and LSTM models [1, 36]. Deep learning models are outperforming traditional methods in discovering hidden patterns and temporal dependencies in large-scale datasets [1, 36]. The combination of a CNN and a BiGRU has achieved up to 97% accuracy in predicting student achievement by providing both local feature extraction and bidirectional temporal modeling [36]. Furthermore, transformer architectures have increased predictive granularity by converting student behaviors into sequential feature vectors through attention mechanisms [1, 8].

D. Explainable Artificial Intelligence (XAI) and Educational Trust

The "black box" nature of models makes it difficult for educators to trust these predictions. In this context, methods such as SHAP and LIME have become widespread to make model decisions more transparent [4, 18, 27]. XAI techniques enable the generation of personalized feedback by revealing which factors influence the prediction and in which direction. In particular, the linguistic rules provided by neuro-fuzzy systems enhance the model's direct interpretability, thereby strengthening educational decision-support processes [18, 22].

E. Indicators Specific to Mathematics Education

Specific research on mathematics education shows that behavioral indicators, such as revisiting resources and resource intensity, are strongly correlated with success [6]. Bakar et al. (2022) suggested incorporating neuroscientific mechanisms such as attention, emotion, and productivity into the AGES model to predict mathematical achievement [15]. Such studies argue that mathematical learning ability can be more accurately predicted not only by academic scores but also by integrating cognitive and psychological processes [37].

F. Research Gaps

A review of the existing literature reveals that the vast majority of studies on the ASSISTments dataset focus either on classical classification-based machine learning algorithms or on information-tracking-based models. These studies mainly address student achievement through sharp class labels, remaining limited in their ability to model uncertainty, incremental progress, and intermediate achievement levels that arise in the student's learning process. In contrast, fuzzy logic-based and neuro-fuzzy approaches have been addressed in only a few studies within the context of the ASSISTments dataset and have not been systematically evaluated.

Although student behavior in online mathematics education inherently involves both continuity and uncertainty, the literature lacks two-stage hybrid approaches that first model this uncertainty as a regression-based achievement score and then proceed to a classification phase for decision support. This highlights the need for a modeling framework that is both pedagogically meaningful and methodologically consistent.

The main gaps in the literature that this study aims to address are summarized below:

- Limited Use of Neuro-Fuzzy Regression Models on the ASSISTments Dataset: The vast majority of studies on the ASSISTments dataset have focused directly on classification using XGBoost, deep learning-based RNN/LSTM models, or knowledge-tracking approaches. However, neuro-fuzzy models like ANFIS can model uncertain learning behaviors more flexibly, as they can predict student success as a continuous score or probability.

While the literature shows that ANFIS provides high accuracy, especially in small and medium-sized datasets, its use as a regression-based predictor in large-scale datasets with time-ordered interaction logs, such as ASSISTments, has been largely neglected. This underscores the necessity of a systematic evaluation of hybrid regression-based neuro-fuzzy models on ASSISTments.

- Treating student achievement solely as a classification problem: A significant portion of current research addresses student performance only as a binary classification problem, ignoring the gradual and uncertain nature of the learning process. However, in online learning environments, student achievement shows gradual increases or decreases over time rather than sudden, sharp transitions.

The literature shows that hybrid models, which first model achievement as a continuous regression outcome and then move it to the classification stage for pedagogical decision support, have not been sufficiently examined on the ASSISTments dataset. This gap necessitates an evaluation of hybrid regression-classification approaches within the context of educational analytics.

- The lack of a clear and in-model approach to addressing uncertainty: In online learning environments, students' interaction speed, problem-solving strategies, and participation levels are inherently fuzzy and variable. However, most existing models treat these behaviors as discrete numerical values, failing to adequately represent situations that require human reasoning, such as a student being "moderately successful" or exhibiting "inconsistent learning behavior."

Although membership functions and linguistic variables in fuzzy logic allow direct modeling of this uncertainty, the use of this approach within hybrid regression and classification frameworks is quite limited in the literature. In particular, the integration of ANFIS as an uncertainty-aware predictor into hybrid structures remains a significant gap in the literature.

- Failure to address pedagogical interpretability and high predictive power together: While powerful ensemble learning methods like RF and XGBoost provide high accuracy, they offer limited transparency for pedagogical decision support systems due to the inability to interpret the internal structure of decision mechanisms directly. Although the literature indicates that XAI methods such as SHAP and LIME address this problem post-prediction, these approaches cannot directly reflect the model's internal decision logic.

In contrast, ANFIS's IF-THEN rule-based structure offers interpretable insights into the learning process. However, studies in the literature that combine this interpretable structure with powerful models such as RF and XGBoost in a hybrid framework are insufficient for the ASSISTments dataset.

- Lack of holistic evaluation of hybrid regression and hybrid classifiers: Existing studies generally focus either solely on regression or solely on classification problems. However, in online mathematics education, student performance needs to be considered both as a continuous indicator of success and as a decision-based classification output.

In the literature, studies that combine ANFIS-based regression outputs with models such as RF and XGBoost to evaluate them within a two-stage hybrid regression-classification architecture are quite limited. This gap necessitates a systematic and comparative examination of hybrid models on the ASSISTments dataset.

- Insufficient modeling of cognitive processes specific to online mathematics education: Mathematics education differs from other disciplines due to its cumulative knowledge structure and high logical dependence. However, most existing models do not address indicators specific to the mathematical problem-solving process, such as hint use, post-error behaviors, cognitive persistence, and learning stability, within a hybrid modeling framework.

This clearly highlights the need for holistic hybrid models that address cognitive and behavioral uncertainties specific to mathematics education at both the regression and classification levels.

In summary, this study addresses complex student interactions in the ASSISTments dataset by combining ANFIS's uncertainty-aware, interpretable regression structure with the highly generalizable capabilities of RF and XGBoost, aiming to fill the gaps identified in the literature from both a hybrid regression and a hybrid classifier perspective. In this respect, the study offers a unique, systematic, and applicable contribution from both methodological and pedagogical standpoints.

III. Materials and Methods

A. Dataset Description

The dataset used in this study was obtained from ASSISTments, an online mathematics learning platform. ASSISTments is an intelligent tutoring system designed to support the problem-solving processes of middle and high school students while collecting detailed, timestamped data on student interactions [38]. The platform integrates teaching and assessment processes by offering students instant feedback, tips, and validation [7, 39, 40].

The ASSISTments dataset records student interactions with math problems in an event-based manner and provides multidimensional behavioral insights for each student-problem interaction [20, 41]. This dataset is widely used, particularly for

student performance prediction, knowledge monitoring, and educational data mining [42]. Furthermore, the fact that ASSISTments data has been used in large-scale data mining competitions such as the KDD Cup 2010 demonstrates that the dataset is an accepted and reliable source within the research community. A sample cross-section of the ASSISTments dataset is given in Table I.

TABLE I. ILLUSTRATIVE SAMPLE FROM THE ASSISTMENTS DATASET

assignment_id	user_id	problem_id	correct	attempt_count	…	skill_id	hint_count	overlap_time
263599	14	93383	0	1		2	2	41131
263599	14	93407	1	1		2	0	29123
263599	14	93400	0	1		2	2	19905
…								

Personality-based features were not used as model inputs; they were evaluated only during student-based data splitting to prevent data leakage. Behavioral features were obtained directly from raw data, while skill-based features were derived to reflect the student's learning progress over time.

The "attempt_count" and "hint_count" attributes reflect students' problem-solving strategies, while "ms_first_response" provides temporal information about engagement and cognitive effort. The "correct" attribute indicates whether the student successfully solved the problem in a given attempt. These attributes allow for the analysis not only of students' outcome-oriented success but also their problem-solving strategies, learning behaviors, and cognitive effort. In particular, "hint_count", "attempt_count", and "ms_first_response" offer significant indicators of students' learning levels and uncertainties in their problem-solving processes [43]. Traditional binary assessment methods are insufficient for modeling these unclear boundaries between learning situations. Managing this uncertainty and predicting student achievement with more humane reasoning constitutes the main motivation for using ANFIS in this study.

B. Data Preprocessing

Educational data obtained from online learning environments inherently contain incomplete records, noisy behavior, and uneven classroom distributions. Therefore, data preprocessing and feature extraction stages are critical for model performance in student performance prediction studies [44]. In this study, a comprehensive data preprocessing and feature engineering process was applied to improve the effectiveness of the proposed hybrid neuro-fuzzy and machine learning model on the ASSISTments dataset.

In the first stage, missing or inconsistent records in the dataset were analyzed. Records lacking "user_id", "assignment_id", and "correct" information were removed from the dataset.

Because the ASSISTments dataset has an event-based structure, the number of interactions among students varies considerably. Some students solve only a few problems, while others have thousands of interactions. This is consistent with the long-tailed distribution frequently reported in the literature [43]. To prevent the model from being dominated by overactive students, student-based (user_id) data splitting and limiting interaction counts were implemented. Furthermore, due to significant variability in student interaction counts within the dataset, students with fewer than 5 interactions were removed, and a maximum interaction limit of 1000 was set for overactive students.

When the "attempt_count" values were analyzed, it was observed that the vast majority were concentrated between 1 and 10. Therefore, in fuzzy category definitions, the maximum number of attempts was set to 10 [7, 43]. Since higher values rarely occur, a threshold has been set at the upper limit, ensuring stable representation and model stability.

The "ms_first_response" property can have excessively large values due to factors such as browser open times and prolonged inactivity. To prevent such values from adding noise to the learning process, the "ms_first_response" property is limited to 600 seconds. Without this limitation, the sensitivity of the fuzzy membership functions decreases, extreme values dominate the model, and the model learns biased rules such as "slow learner = fail." This approach is consistent with clipping strategies proposed in the literature for time-based learning variables [43, 44].

"opportunity_count" is a pedagogically significant attribute representing how many times a student has encountered a particular skill. Therefore, it has not been directly filtered but has been included in the model to reflect the learning process.

Pre-cleaning analyses show that the distribution of student interactions is severely imbalanced, with a small number of students dominating the dataset. After applying cleaning and preprocessing criteria, the number of interactions per student showed a more balanced distribution, with extreme values eliminated. Furthermore, normalization was performed to improve the stability of the fuzzy logic-based model's membership functions and to ensure a fair comparison across machine learning algorithms [45]. Min-max normalization was applied to "attempt_count" and "hint_count", robust normalization to "ms_response_time" and "overlap_time", and log+min-max normalization to "opportunity_count". This helped train the proposed hybrid model more stably and generalizable.

C. Feature Engineering

In this study, to more accurately model students' online mathematics learning behaviors, the features obtained from the raw dataset were normalized, and behavioral and performance-based derived features commonly used in the literature were created.

The raw features in the dataset capture students' core interactions during the problem-solving process. While these

raw features directly reflect students' problem-solving behaviors, their direct use as model input is limited by scale differences and individual variation. Instead of using raw features directly, derived features were created to better represent students' performance and behavioral tendencies during the learning process. These features were designed to reflect the student's cognitive effort and learning status in the problem-solving process [43]. In this context, the following features have been derived: Prior Correct Ratio (PCR), Prior Attempt Mean (PAM), Prior Hint Mean (PHM), Recent Performance Trend (RPT), Skill Difficulty Index (SDI), and Student Experience Level (SEL).

PCR is a ratio-based trait that reflects a student's level of success in previous interactions with a specific problem or skill [6, 46]. The PCR value is in the range of [0,1] and does not require normalization.

$$\mathrm{PCR}_{u,s} = \frac{\sum_{i=1}^{n} Correct_{u,s}^{(i)}}{n} \quad (1)$$

Here, u is a student, s is a problem or a skill, and n is the number of previous attempts.

PAM is a behavioral indicator that represents the average number of attempts a student makes for problems they have encountered in the past. This characteristic reflects the student's mathematical cognitive persistence and learning speed [6], [46], [47]. Since the PAM value is not in the [0,1] range, pre-model min-max normalization should be applied. A high PAM value indicates that the student generally needs more trials.

$$\mathrm{PAM_u} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{attempt_count_u^{(i)}} \quad (2)$$

Here, u is a student, and n is the number of problems previously solved.

PHM represents the average number of hints a student uses across all problem-solving steps before the current interaction. This characteristic is a historical indicator of a student's help-seeking behavior. Hint usage carries a vague signal about how well a student understands the subject, and overuse is associated with an increased risk of failure [38].

$$\mathrm{PHM}_i = \frac{1}{N_i}\sum_{j=1}^{i-1} hint_coun\,t_j \quad (3)$$

Here, i is the student's time-ordered current interaction index, $N_i = i - 1$ is the total number of interactions before the current interaction, $hint_count_j$ is the number of hints the student used in the j[th] interaction.

RPT captures short-term learning dynamics, information fluctuations, and learning trends, which are particularly important in adaptive learning environments. Given the cumulative nature of mathematics education, short-term dips are considered a signal for early intervention [8], [48].

For the last k interactions (3-5):

$$\mathrm{RPT}_i = \frac{1}{k}\sum_{j=i-k}^{i-1} correc\,t_j \quad (4)$$

Here, i is the student's i^{th} problem-solving step.

SDI is the overall difficulty level of a given mathematical skill across all students. This metric allows the model to make fair and balanced assessments by normalizing individual performance with aggregate data [49], [50], [51].

$$\mathrm{SDI}_s = 1 - \frac{1}{M}\sum_{j=1}^{M} Correc\,t_s^{(j)} \quad (5)$$

Here, s is the skill, and M is the total number of trials for that skill.

SDI is calculated at the skill level, independent of the problem, and provides a common difficulty indicator for all students (SDI → 0: Easy skill, SDI → 1: Difficult skill).

SEL is a composite pedagogical indicator representing a student's overall experience and proficiency level in an online mathematics learning environment. This characteristic translates the student's familiarity with the system and resource-intensive usage habits into the model [15]. SEL is not a variable with an implicit formula. It is defined as follows:

$$\mathrm{SEL}_i = F(\mathrm{PCR}_i, \mathrm{PAM}_i, \mathrm{PHM}_i, \mathrm{OC}_i) \quad (6)$$

Here, i is the student's time-ordered current interaction index, and $F(\cdot)$ is the Fuzzy Inference Mechanism (ANFIS).

The derived features capture both past performance and real-time behavioral patterns, providing a richer representation of student learning dynamics. All features used in the study are listed in Table II.

TABLE II. CHARACTERISTICS USED IN THE STUDY

Feature	Reason
attempt_count (AC)	The solution reflects the intensity of the behavior; in ANFIS, it is meaningful in terms of "Low–Medium–High" memberships.
hint_count (HC)	Help-seeking behavior can be interpreted pedagogically using vague rules.
ms_response_time (RT)	ML models can tolerate high-dimensional correlation; this increases rule complexity in ANFIS.
Prior Correct Ratio (PCR)	The student's past achievements are critical to their learning.
Prior Attempt Mean (PAM)	Long-term effort level: a stable statistical characteristic.
Prior Hint Mean (PHM)	The constant need for help reflects the learning strategy.
Recent Performance Trend (RPT)	The short-term learning tendency is suitable for trend-based rules in ANFIS.
opportunity_count (OC)	The number of times a student encounters a skill is a key context for both statistical and rule-based models.
overlap_time (OT)	Student waiting time for the question
Skill Difficulty Index (SDI)	Content difficulty: pedagogically meaningful within ANFIS guidelines.
Student Experience Level (SEL)	Cumulative learning level: intuitive linguistic variable in ANFIS.

D. Inter-Feature Relationship Analysis

As part of the feature identification process, a correlation analysis was performed to assess relationships among the extracted features. The resulting Pearson correlation matrix, shown in Table III, provides insight into potential feature dependencies before neuro-fuzzy logic and machine learning

modeling. The analysis included both raw and derived numerical features, excluding descriptive variables.

The vast majority of correlation coefficients are at the $|r|<0.40$ level. This indicates that there is no high-level linear dependence between the features and that the risk of multicollinearity is generally low. A moderate positive correlation ($r=0.46$) was observed between RT and OT. This relationship shows that time overlap increases as the student's first response time increases. There is a relatively high positive correlation ($r = 0.58$) between HC and PHM. This indicates that the student's past habit of using hints is reflected in their current problem behavior. PCR shows a strong negative relationship with both PAM ($r = -0.53$) and PHM ($r = -0.62$). This finding reveals that students with high past success tend to try less and use fewer hints. The correlation between the RPT feature and all other features is negligible, suggesting that the variable provides an independent and complementary source of information for the model. The very low correlation of OC with other characteristics suggests that this variable is largely independent of other behavioral traits as a time-based measure.

TABLE III. PEARSON CORRELATION MATRIX FOR THE FEATURE SET

Feature	AC	HC	RT	OT	SDI	SEL	RPT	PCR	PAM	PHM	OC
AC	1.00	0.38	0.04	0.17	0.11	-0.04	-0.02	-0.19	0.27	0.16	0.00
HC	0.38	1.00	0.00	0.10	0.22	-0.09	-0.03	-0.37	0.19	**0.58**	-0.01
RT	0.04	0.00	1.00	**0.46**	0.09	0.03	0.02	0.02	0.02	0.00	0.03
OT	0.17	0.10	**0.46**	1.00	0.07	0.01	0.01	-0.03	0.05	0.04	0.01
SDI	0.11	0.22	0.09	0.07	1.00	0.03	0.01	-0.35	0.21	0.31	-0.06
SEL	-0.04	-0.09	0.03	0.01	0.03	1.00	0.01	0.13	-0.08	-0.11	0.16
RPT	-0.02	-0.03	0.02	0.01	0.01	0.01	1.00	-0.01	0.01	0.01	0.01
PCR	-0.19	-0.37	0.02	-0.03	-0.35	0.13	-0.01	1.00	**-0.53**	**-0.62**	0.08
PAM	0.27	0.19	0.02	0.05	0.21	-0.08	0.01	**-0.53**	1.00	**0.41**	-0.04
PHM	0.16	**0.58**	0.00	0.04	0.31	-0.11	0.01	**-0.62**	**0.41**	1.00	-0.04
OC	0.00	-0.01	0.03	0.01	-0.06	0.16	0.01	0.08	-0.04	-0.04	1.00

Overall, the correlation analysis results indicate that including the selected features in both neurofuzzy and machine learning models is statistically feasible and that there is no significant multicollinearity.

In addition to correlation analysis, feature redundancy was further examined using the Variance Inflation Factor (VIF). VIF quantifies the degree of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to linear dependence among predictors. The VIF analysis was conducted on all numerical raw and derived features extracted from the ASSISTments dataset. Following established guidelines, a VIF threshold of 10 was adopted, where values below this limit indicate acceptable levels of multicollinearity (Table IV).

TABLE IV. VIF TABLE

Feature	VIF	Comment
AC	3.79	Moderate correlation - acceptable.
HC	2.03	Low multi-connectivity
RT	1.96	Low multi-connectivity
OT	1.49	Very low multi-connectivity
SDI	**8.04**	**High – caution advised**
SEL	1.98	Low multi-connectivity
RPT	1.01	Independent variable
PCR	**4.26**	Medium-high
PAM	**5.84**	**High on the border**
PHM	2.78	Middle
OC	2.21	Low

VIF was calculated to quantify the degree of multicollinearity among the independent variables. Generally, VIF values <5 are acceptable; values between 5 and 10 require attention; and values ≥ 10 indicate significant multicollinearity. According to the analysis results, the vast majority of features have VIF<5. This indicates that the features contribute largely independently to the model. The RPT feature, with a VIF$\approx$1, is an almost completely independent source of information and a strong complement to the model. AC, HC, RT, OT, and SEL have low-to-medium VIF values and do not pose a risk in terms of linear dependence. PAM (VIF$\approx$5.84) and especially SDI (VIF$\approx$8.04) exhibit relatively high VIF values. This suggests that these variables share common variance with some other inputs in the model. However, since these values are below the critical threshold of 10, they do not warrant direct exclusion. These results indicate that the feature set is generally well-balanced and does not exhibit significant multicollinearity. When the VIF results are evaluated together with the Pearson correlation matrix, it becomes clear that the relatively high VIF values observed are due to natural overlap.

The SDI feature shows moderate negative correlations with PCR and moderate positive correlations with PHM and PAM. These relationships stem from the fact that problem difficulty is naturally linked to the student's previous success, trial, and hint-use behaviors. Therefore, the high VIF value is not a statistical problem, but a reflection of a pedagogically significant overlap.

There is a moderate-to-high negative correlation between PAM and PCR features. It is expected that a student's experience of more trials is generally associated with a lower success rate. The increase in VIF values is a mathematical consequence of

this behavioral correlation and does not mean the model is meaningless.

There is a significant positive relationship between HC and PHM. The relatively higher VIF values result from the inclusion of both short- and long-term representations of the same concept in the model. This requires attention, especially in linear models, but does not pose a problem for nonlinear models.

Features such as RPT, OT, and RT exhibit low correlation and low VIF values. These variables provide independent and complementary information to the model.

No feature exceeds the critical multicollinearity threshold (VIF≥10). Therefore, all features are preserved, especially in ANFIS and machine learning-based nonlinear models.

E. Problem Formulation

In online learning environments, students' problem-solving performance exhibits a nonlinear, complex structure due to individual learning differences, prior knowledge levels, problem difficulty, and behavioral uncertainties arising during interaction. This uncertainty makes it difficult to predict student success with both high accuracy and pedagogical interpretation. Therefore, hybrid prediction mechanisms that can explicitly model uncertainty in student performance and offer high predictive power are needed in online mathematics education.

The ASSISTments dataset consists of multidimensional observations derived from students' time-ordered interactions, each reflecting the student's cognitive and behavioral state. Each student interaction is defined as follows:

$$\mathrm{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}] \in \mathbb{R}^n \tag{7}$$

Here, x_i represents raw and statistical features derived from the student's past performance, problem characteristics, and learning process.

In the ASSISTments dataset, the output variable is defined as a binary label containing correct or incorrect response information for each student-problem interaction:

$$y_i \in \{0,1\} \tag{8}$$

However, the Subtractive Clustering-based ANFIS model used in this study inherently produces a continuous-valued output due to its Sugeno-type fuzzy inference mechanism. Therefore, in the first stage of the study, the problem was treated as a hybrid regression problem, and the model output was defined as follows:

$$\hat{y}_i^{(m)} = f^{(m)}(\mathrm{x}_i), m \in \{\text{ANFIS}, RF, \text{XGBoost}\} \tag{9}$$

Here, $\hat{y}_i^{(m)}$ is interpreted as a continuous value corresponding to the student's success score on the given problem or the probability of giving the correct answer.

In this context, ANFIS is an interpretable regression estimator that models ambiguous student behavior using fuzzy rules, whereas RF and XGBoost are robust regression models with high capacity to capture nonlinear patterns. All models were trained and evaluated to predict the same continuous target variable, ensuring methodological consistency.

While regression-based predictions offer rich information about the probability of student success, decision-based outcomes are frequently needed in educational applications. Therefore, in the second phase of the study, the problem was reformulated as a hybrid classification problem. In this phase, using the continuous outputs obtained from the regression models, the final class label was defined as follows:

$$\tilde{y}_i = \begin{cases} 1, & \hat{y}_i \geq \tau \\ 0, & \hat{y}_i < \tau \end{cases} \tag{10}$$

Here, τ represents the decision threshold set on the validation data.

In this approach, ANFIS acts as a pre-predictor, providing fuzzy-logic-based intermediate outputs to the decision-making process, while RF and XGBoost serve as higher-level classifiers that make the final decision based on these outputs. This structure offers a unique hybrid classification framework that combines both uncertainty modeling and high classification accuracy under one roof.

F. ANFIS Model and Learning Process

In the field of educational data mining, student knowledge levels, learning speeds, and performance changes often involve ambiguous transitions rather than precise boundaries; therefore, ANFIS offers a suitable and powerful framework for modeling such patterns [42].

In the generated ANFIS model, the number of membership functions and rules is automatically determined based on data density using the subtractive clustering method. Sugeno-type output is used. The model's training parameters are defined as Epoch=100, StepSize=0.01, and ErrorGoal=0. In the subtractive clustering method, the cluster radius parameter is a critical hyperparameter that directly determines the complexity of the ANFIS model. Small radius values result in more clusters in the data space and, consequently, a high number of fuzzy rules, while large radius values lead to oversimplification of the model. With a cluster radius of 0.5, an average of 5–8 membership functions was obtained for each input variable.

ANFIS is a model based on the supervised learning paradigm, which requires the initial definition of the target variable during training. In the subtractive clustering method, the fuzzy rule parameters are initially estimated using the least-squares method. At this stage, the prior parameters are assumed to be constant, and the rule firing strengths are calculated for each training instance [34, 45]. The normalized firing strength for each rule is defined as follows:

$$\bar{w}_k = \frac{w_k}{\sum_{l=1}^{K} w_l} \tag{11}$$

Here, w_k represents the firing strength of the kth fuzzy rule.

The model output is calculated as follows, using normalized firing strengths:

$$\hat{y} = \sum_{k=1}^{K} \bar{w}_k \left(p_{k0} + \sum_{j=1}^{n} p_{kj}\, x_j \right) \tag{12}$$

When this expression is rearranged in terms of linear parameters, a classical linear regression problem is obtained. This linear system is solved using the least-squares method,

considering all training examples, and the initial output parameters $\{p_{k0}, p_{k1}, \ldots, p_{kn}\}$ are computed for each rule.

After determining the initial output parameters, the ANFIS model was trained using a hybrid learning algorithm. In this algorithm, during the forward traversal phase, the prior parameters are held constant, and the output parameters are updated using the least-squares method. In the backpropagation phase, the output parameters are treated as constants, and the prior parameters are optimized using gradient descent. This two-stage learning process increases both the accuracy and generalization ability of the model [45].

G. Random Forest and Learning Process

In this study, the RF model was used in regression mode to predict a continuous-valued student achievement score from pairwise-labeled student interactions. RF generates the final prediction by averaging the outputs of multiple decision trees created using bootstrap sampling [52]. This structure offers significant advantages, including the ability to model nonlinear relationships and resilience against overlearning. Each decision tree learns the relationship between the input features and the target variable to produce a continuously valuable output. The RF output is calculated as follows:

$$\hat{y}_i^{RF} = \frac{1}{T}\sum_{t=1}^{T} f_t\,(\mathbf{x}_i) \tag{13}$$

Here, T represents the number of trees, and $f_t(\cdot)$ represents the t[th] decision tree [52].

This structure allows RF to generate predictions in the same target space as the continuous output produced by the ANFIS model, enabling a fair comparison.

H. XGBoost and Learning Process

XGBoost is a regularized decision tree ensemble method based on the gradient boosting principle. The model improves its performance by learning a new tree in each iteration, focusing on minimizing the errors of previous predictions [52]. In this study, the XGBoost method is used as a regression method to estimate continuous student achievement scores. The general form of the XGBoost model can be expressed as follows:

$$\hat{y}_i^{XGB} = \sum_{m=1}^{M} \gamma_m\, h_m(\mathrm{x}_i) \tag{14}$$

Here, $h_m(\cdot)$ represents the m^{th} regression tree, and γ_m represents the learning rate [53].

This structure allows XGBoost to generate predictions in the same target space as the continuous outputs of other models, enabling a fair comparison.

I. Proposed Hybrid Model: ANFIS–RF–XGBoost-Based Student Achievement Prediction Framework

In this study, the proposed architecture comprises two complementary learning stages: regression-based pre-prediction with uncertainty awareness and classification-based final output generation, both focused on decision support. Fig. 1 presents the general flowchart of the hybrid architecture proposed in this study.

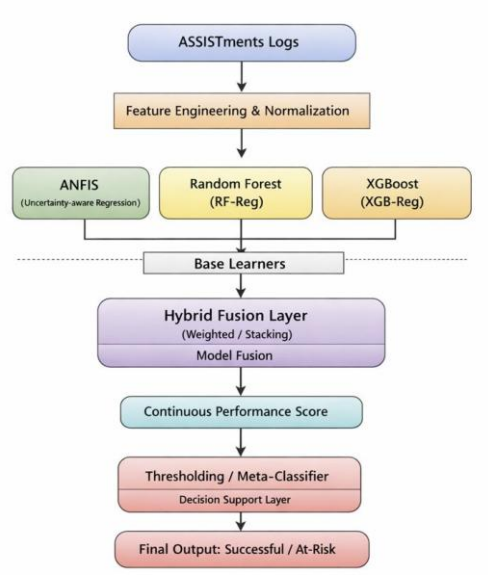


Fig. 1. Proposed hybrid model.

The raw interaction logs in the ASSISTments dataset reflect students' problem-solving behaviors in chronological order. The feature engineering process performed on the dataset is described in detail in Section III-C.

In the Base Learners phase, ANFIS, RF, and XGBoost regression processes are performed. With ANFIS, student behaviors are represented by linguistic, interpretable rules, and a continuously valued achievement score is obtained. In addition to ANFIS, XGBoost models were trained in regression mode to capture nonlinear patterns with high accuracy, and RF models were trained for robustness to noise and stable prediction. These models complement ANFIS by enabling learning of complex feature interactions, robustness to noise, and high generalization performance.

Three independent models are trained using the same input properties x_i

$$\begin{aligned} \hat{y}_i^{ANFIS} &= f_{ANFIS}(\mathrm{x}_i) \\ \hat{y}_i^{RF} &= f_{RF}(\mathrm{x}_i) \\ \hat{y}_i^{XGB} &= f_{XGBoost}(\mathrm{x}_i) \end{aligned} \tag{15}$$

All of these models are structured in regression mode to predict a continuous student achievement score from binary-labeled data.

In the next stage, the regression-based predictions are fed into the fusion layer, which is at the heart of the hybrid architecture. Two different hybridization strategies, weighted ensemble and stacking, were evaluated in this study.

In the weighted ensemble approach, the output of each model is weighted according to its performance on the validation data:

$$\hat{y}_i^{Hybrid} = \alpha\hat{y}_i^{ANFIS} + \beta\hat{y}_i^{RF} + \gamma\hat{y}_i^{XGB} \tag{16}$$

Here, $\alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \geq 0$, and this can be determined based on validation set performance, model reliability, or grid search optimization. Thus, it balances the weaknesses of individual models and increases the stability of the regression output.

In stacking-based hybrid learning, the ANFIS, RF, and XGBoost outputs are treated as meta-features and fed to a second-level learner. This meta-learner is structured as a linear regression classifier for regression and a logistic regression classifier for classification. This stacking structure produces a more powerful prediction by learning complementary information between models.

$$z_i = [\hat{y}_i^{ANFIS}, \hat{y}_i^{RF}, \hat{y}_i^{XGB}] \quad (17)$$

Here, z_i is the meta-input vector.

For the final estimate:

$$\hat{y}_i^{Hybrid} = g(z_i) \quad (18)$$

Here, $g(\cdot)$ is chosen to be either a linear or a low-complexity regression model.

The consistent success scores were transformed into a binary classification problem for decision support in educational applications. This transformation was performed using the following thresholding function:

$$\tilde{y}_i = \begin{cases} 1, & \hat{y}_i^{\text{hybrid}} \geq \tau \\ 0, & \hat{y}_i^{\text{hybrid}} < \tau \end{cases} \quad (19)$$

Here, the threshold value τ is optimized on the validation set.

The proposed hybrid architecture offers the following advantages:

- Uncertainty Awareness: ANFIS explicitly models fuzzy patterns in student behavior.
- High Predictive Power: RF and XGBoost successfully capture complex and nonlinear relationships.
- Interpretability: ANFIS rules allow for pedagogical interpretation of prediction results.
- Flexible Use: The architecture is naturally adaptable to both regression and classification scenarios.

J. Performance Evaluation Metrics

The proposed hybrid architecture treats student achievement as both a continuous performance score (regression) and a binary decision outcome (classification). Therefore, model performance was analyzed using a two-stage evaluation strategy appropriate to the nature of the problem. In the first stage, the accuracy of regression-based predictions was measured, and in the second stage, the classification performance of the resulting hybrid outcomes was evaluated.

This multi-level performance evaluation strategy comprehensively demonstrates not only the prediction accuracy of the proposed hybrid model but also its practical usability for early warning and pedagogical decision-support systems.

1) Regression-based evaluation metrics: The ANFIS, RF, and XGBoost models were structured in regression mode to predict the student's consistent success score on the relevant problem. The metrics used at this stage measure the level of error between the predicted values and the actual labels. For regression-based models, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) metrics were used [11, 22, 24].

Furthermore, the continuous predictions generated by regression models are subsequently converted to a binary classification problem for evaluating decision performance. Fixed threshold values are commonly used for this process. However, the literature indicates that such fixed thresholds do not guarantee optimal decision performance, especially when the class distribution is unbalanced and false-positive and false-negative errors have asymmetric costs [54, 55]. To overcome this limitation, an adaptive thresholding method based on the ROC curve was used in this study, rather than a fixed thresholding approach. The adaptive threshold was set to the point on the ROC curve that best balances sensitivity and specificity. For this purpose, the Youden Index (J), widely used in the literature, was adopted [56]. The Youden index is defined as follows:

$$J = TPR - FPR \quad (20)$$

This criterion determines the decision boundary that best distinguishes between classes by selecting a threshold value that maximizes the true positive rate while minimizing the false positive rate.

The ROC-based adaptive thresholding approach allows evaluation not only of the continuous predictive accuracy of models but also of their decision-making success, in a manner sensitive to data distribution. Particularly in online mathematics learning environments, considering the pedagogical consequences of incorrectly classifying a failing student as successful, this approach offers a more realistic and reliable assessment framework [44, 57].

2) Classification-based evaluation metrics: Regression-based hybrid outputs were converted into a binary classification problem using a threshold or a meta-classifier for decision support. Classification performance was evaluated using accuracy, recall, precision, and F1-score [4, 47, 58].

IV. Experimental Results

A. Experimental Setup

This section describes the experimental setup used to objectively and reproducibly evaluate the performance of the proposed hybrid neuro-fuzzy and machine-learning-based architecture. The experiments were designed within a multi-stage framework aiming to predict student achievement levels in online mathematics learning environments in both continuous (regression) and binary (classification) forms. In this context, the comparative performance of single learners and hybrid regression and hybrid classifier models was analyzed on raw and derived feature sets using the ASSISTments dataset. The entire experimental process was structured with the principles of fair comparison and data leakage prevention in mind.

B. Dataset and Splitting Strategy

In the experimental analyses, a student-based splitting strategy was adopted to prevent data leakage and to realistically evaluate the models' generalization ability. In this approach, the interactions of the same student were included in only one subset (training, validation, or testing). Thus, by measuring the model's performance on students it has not previously encountered, an evaluation closer to practical use cases was provided. The dataset was divided into three subsets: 60% for training, 20% for validation, and 20% for testing. The training set was used to learn the model parameters, the validation set for hyperparameter adjustments and determining hybrid model weights, and the testing set for the final performance evaluation. In the classification phase, the distribution of success and failure labels was analyzed, focusing on recall and F1-score to mitigate class imbalance and prevent skewed performance metrics. Thanks to this splitting strategy, the reliability of the proposed hybrid regression and hybrid classifier models in terms of both prediction accuracy and early warning systems was evaluated consistently and fairly.

C. Experimental Scenarios

To comprehensively evaluate the effectiveness of the proposed architecture, numerous experimental scenarios were designed that incorporated different model types, feature sets, and learning paradigms. These experimental scenarios aim to analyze the performance of individual models and clearly demonstrate the contribution of hybrid structures. In this context, the experiments were grouped into three main categories: individual regression models, hybrid regression models, and hybrid classification models.

In individual regression models, student achievement was treated as a continuous performance score, and ANFIS, RF, and XGBoost models were trained in regression mode. The goal was to compare the fundamental predictive capabilities of different learning paradigms on the ASSISTments dataset. In this context, analysis was performed with three regression models and three feature sets (raw, derived, and hybrid). These scenarios enable analysis of the impact of feature engineering and model structure on regression performance.

Hybrid regression scenarios aim to achieve more stable, accurate predictions by combining outputs from individual models. In this stage, ANFIS, RF, and XGBoost regression outputs were used as inputs to hybrid coupling mechanisms. These experiments were designed to evaluate the performance change and generalization status of the hybrid approach compared to single models.

In the hybrid classification scenario, the continuous success scores from the hybrid regression models were converted into binary classification labels for decision support. In this scenario, students were divided into two classes. This scenario aims to evaluate the applicability of the proposed architecture for early warning systems and its potential for pedagogical decision support.

D. Comparison of Individual Models

In this section, the performance of the individual regression models that constitute the core components of the proposed hybrid architecture is comprehensively analyzed. Student achievement is treated as a continuous output variable, and the ANFIS, RF, and XGBoost models are evaluated within the regression framework. Each model was trained and tested separately using raw features, derived features, and hybrid feature sets to examine the effect of feature representations on prediction performance. The test results obtained after nine analyses are given in Table V.

TABLE V. PERFORMANCE OF SINGLE REGRESSION MODELS

	Regression				Decision				
Model	**Features**	**RMSE**	**MAE**	**R^2**	**Threshold**	**Accuracy**	**Precision**	**Recall**	**F1**
ANFIS	Raw	0.2633	0.1409	0.6699	0.4825	0.922063	0.902544	0.9962	0.9471
	Derived	0.4159	0.3461	0.1765	0.7169	0.681849	0.828835	0.6874	0.7515
	Hybrid	0.2545	0.1411	0.6917	0.6971	0.916655	0.918273	0.9670	0.9420
RF	Raw	0.2561	0.1352	0.6878	0.8174	0.916943	0.910889	0.9769	0.9427
	Derived	0.4082	0.3301	0.2065	0.6790	0.714277	0.821494	0.7561	0.7874
	Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
XGBoost	Raw	0.2597	0.1386	0.6790	0.8058	0.918995	0.906226	0.9863	0.9446
	Derived	0.4109	0.3399	0.1961	0.6966	0.704823	0.824742	0.7343	0.7769
	Hybrid	0.2434	0.1261	0.7179	0.6899	0.914872	0.927508	0.9528	0.9400

The table's results point to three clear and consistent findings. There is a significant performance decrease in all models when derived features are used alone. The hybrid feature set (Raw + Derived) is the most balanced and robust structure for all models. RF and XGBoost perform better in regression than ANFIS; however, ANFIS remains competitive in decision-making accuracy.

The lowest RMSE was obtained with the RF-Hybrid (0.2369), and the lowest MAE was obtained with the RF-Hybrid (0.1146). This result shows that RF is the most successful model in the hybrid feature space, with both lower mean error and lower extreme sensitivity. ANFIS-Hybrid (RMSE = 0.2545) and XGB-Hybrid (RMSE = 0.2434) are robust; however, RF-Hybrid is significantly ahead.

The highest R^2 values were obtained in the RF-Hybrid (0.7327) model, followed by the XGBoost-Hybrid (0.7179) and ANFIS-Hybrid (0.6917) models. This clearly shows that hybrid features enable the model to better explain the variance in the output variable. In particular, the fact that derived features alone dramatically reduce R^2 (≈0.17–0.20) indicates that they do not replace raw information but rather play a complementary role.

The highest accuracy was obtained with the ANFIS-Raw (0.9221) model, followed by the XGBoost-Raw (0.9190) and RF-Raw (0.9169) models. It is worth noting that raw features yield clearer classification signals when using decision thresholds.

The highest recall value was obtained in the ANFIS – Raw (0.9962) model, and the most balanced precision-recall value was obtained in the RF – Hybrid (0.9365 / 0.9382) model. The ANFIS-Raw configuration exhibits a recall-oriented character, while the RF-Hybrid model shows a balance-oriented character.

The highest F1 value was obtained in the ANFIS – Raw (0.9471) model, followed by the closely related XGBoost – Raw (0.9446) and RF – Hybrid (0.9373) models. This indicates that raw features are highly effective for classification, especially with ANFIS and XGB.

The findings show that using derived features alone reduces both regression and decision success, whereas the hybrid structure, which combines raw and derived features, provides significant and consistent improvements, particularly in regression performance. In terms of classification, raw features were observed to contain more distinct discriminative information, especially in the ANFIS and XGBoost models.

E. Hybrid Regression Results

1) Weighted ensemble performance: Ensemble learning approaches aim to reduce the generalization errors of individual models by combining their outputs, leveraging models with different assumptions and learning mechanisms [59], [60]. Weighted ensemble methods, in particular, offer a more balanced and effective estimation mechanism by accounting for each sub-model's validation performance, whereas simple averaging approaches assume all models contribute equally [60]. In this study, the uncertainty modeling and interpretability advantages of ANFIS are combined with the high generalization capabilities of RF and XGBoost models, and the contributions of the sub-models to the ensemble output are adaptively determined based on their validation-set error. Optimizing the weights inversely proportional to the RMSE values calculated on the validation data ensures that models with lower prediction errors have a greater impact on the final decision. This approach is consistent with the performance-based weighting strategies proposed in the literature [60], [61] and aims to contribute to more reliable prediction of student success in online mathematics education, in terms of both accuracy and stability. The most successful validation results, including RMSE-based weighted ensemble analysis and single regression, are shown in Table VI.

TABLE VI. WEIGHTED HYBRID VS. SINGLE MODELS

Model	RMSE	MAE	R^2	Threshold	Accuracy	Precision	Recall	F1
RF+Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
Weighted Hybrid	0.2414	0.1255	0.7226	0.7014	0.916732	0.928630	0.9544	0.9413

Table VI evaluates the regression-based prediction performance and the classification performance derived from these predictions. The RF+Hybrid model has lower RMSE (0.2369) and MAE (0.1146) than the weighted hybrid model. However, its R^2 (0.7327) value is higher. These findings indicate that the RF-based hybrid approach has higher explanatory power and a lower error rate in predicting continuous success scores. Therefore, the RF+Hybrid model is more advantageous in terms of pure regression accuracy.

In contrast, when classification performance is examined, the weighted hybrid model offers more balanced and robust results. Although the RMSE and MAE values of the weighted hybrid model are slightly higher than those of the RF+Hybrid, it is superior in terms of accuracy (0.9167), recall (0.9544), and, especially, the F1 score (0.9413). This increase in recall value is critical, particularly in avoiding missing the "successful student" class, and yields a more pedagogically meaningful result in the context of educational analytics.

When the threshold values are examined, it is seen that the RF+Hybrid model uses a higher threshold (0.7135), whereas the weighted hybrid model uses a lower one (0.7014). This indicates that the weighted hybrid model creates a more comprehensive decision threshold, thereby increasing true positives and positively impacting Recall and F1 scores.

In conclusion, while the RF+Hybrid model provides lower error and higher explanatory power for continuous success prediction, the weighted hybrid model offers a more appropriate, balanced performance in classification-based decision-support scenarios—especially in online learning environments that require early risk detection and intervention. This finding clearly demonstrates that, in hybrid learning analytics studies that consider both regression and classification objectives, model selection should be guided by the problem context.

2) Stacking-based hybrid performance: Stacking (stacked generalization) is an advanced ensemble learning approach that combines the outputs of multiple base learners via a higher-level meta-learner and is particularly notable for its ability to effectively integrate the complementary strengths of different model types [62, 63]. In this approach, instead of using the predictions generated by the baseline models directly as the final decision, they are reweighted by a model learned in the second stage, thus incorporating nonlinear relationships between model outputs into the learning process [60]. In this

study, the outputs of ANFIS, RF, and XGBoost models were combined within a stacking framework. A linear regression model, which is simple and limits overfitting, was chosen as the meta-learner, and the final student success prediction was made using the continuous success scores from the baseline models. This structure is consistent with studies in the literature showing that heterogeneous ensemble models offer higher prediction accuracy compared to singular and weighted ensemble approaches [61, 63] and aims to more effectively model complex and uncertain student interactions in online mathematics education. The results of the hybrid regression analysis and the most successful singular regression results are given in Table VII.

TABLE VII. HYBRID REGRESSION AND SINGLE-DOMAIN REGRESSION RESULTS

Model	RMSE	MAE	R^2	Threshold	Accuracy	Precision	Recall	F1
RF+Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
Weighted Hybrid	0.2414	0.1255	0.7226	0.7014	0.916732	0.928630	0.9544	0.9413
Stacking Hybrid	0.2503	0.1280	0.7017	0.5801	0.916157	0.929559	0.9524	0.9408

The results show that different hybridization strategies achieve a significant balance between minimizing error and maintaining decision sensitivity. In terms of regression performance, the RF+Hybrid model exhibits the lowest RMSE (0.2369) and MAE (0.1146), while also offering the highest R^2 (0.7327). This indicates that the RF-based hybrid structure produces more accurate continuous predictions of student achievement scores by effectively modeling nonlinear relationships. Therefore, the RF+Hybrid approach stands out in scenarios where pure regression accuracy and error minimization are prioritized. In contrast, while weighted and stacking hybrid models exhibit relatively higher regression errors, they offer more balanced and competitive classification performance. Specifically, the weighted hybrid model achieves the highest recall (0.9544) and F1 score (0.9413), demonstrating a more sensitive structure for accurately identifying at-risk students. Similarly, the stacking hybrid model, despite using a lower decision threshold (0.5801), demonstrated comparable recall (0.9524) and F1 (0.9408) scores. This finding shows that the meta-learning-based stacking approach can leverage the complementary information content of different regressors.

When examining the decision thresholds, it is observed that the RF+Hybrid model adopts a more conservative classification strategy with a higher threshold (0.7135), whereas the weighted hybrid and, especially, the stacking hybrid models increase the true positive rate by operating with lower thresholds. This preference directly affected the precision-recall trade-off, resulting in higher recall and F1 scores.

Overall, the RF+Hybrid model offers the strongest performance in continuous success prediction, while the weighted and stacking hybrid models are better suited for classification-based decision-support applications. These results clearly demonstrate that in online education analytics, rather than a single "best" model, different hybrid regression strategies should be preferred depending on the problem objectives (error minimization vs. early risk detection).

F. Meta-Classification Results

This section presents the results obtained by combining continuous success scores derived from regression models within weighted ensemble and stacking-based hybrid classification structures. The analyses show that hybrid approaches, while not significantly improving classification accuracy over single models, produce more stable and consistent results across different experimental scenarios and feature sets. This supports the usability of hybrid architectures as reliable early warning systems in online mathematics education.

In the proposed meta-classification structure, raw or derived features were not directly fed to the classifier; instead, continuous success scores obtained from ANFIS, RF, and XGBoost regression models were used as meta-features. Thus, different learning biases of the base models were integrated at the meta-level to arrive at the final decision. Classification was performed using logistic regression. After analysis, the accuracy was 0.916560, the precision was 0.929140, the recall was 0.9535, and the F1 score was 0.9412.

When the experimental results are examined, it is generally observed that the individual models and the proposed hybrid structures exhibit similar performance on metrics such as accuracy and F1-score. This proves that the pedagogical variables (PCR, PAM, RPT, etc.) derived from raw data during the feature engineering phase have a very strong signal capacity in representing the fundamental patterns in student behavior. Therefore, the similarity in performance between the models is considered a saturation point resulting from the high quality of the distinguishing features in the dataset.

G. Interpretability Analysis (ANFIS)

This section outlines the interpretability advantages of the ANFIS model within the proposed hybrid architecture. In online mathematics learning environments, high prediction accuracy is as critical as the pedagogical interpretability of these predictions. Unlike black-box machine learning models, ANFIS's Sugeno-type fuzzy inference mechanism allows for a clear examination of the underlying logical relationships in predicting success by representing student interactions through "If-Then" rules. In this context, the ANFIS model is considered not only a predictor within the hybrid structure but also an explanatory component that transforms the uncertainty in student behavior into a pedagogically interpretable form. Three rules have been established in the educational process using the hybrid feature set and the ANFIS model. These are:

Rule 1: IF AC is low [MF 1] AND HC is low [MF 1] AND RT is low [MF 1] AND OT is low [MF 1] AND SDI is low [MF 1] AND SEL is low [MF 1] AND RPT is high [MF 1] AND PCR is low [MF 1] AND PAM is low [MF 1] AND PHM is low [MF 1] AND OC is low [MF 1] THEN output = -24.596*x -6.812*x

-0.123*x -12.608*x -0.436*x +0.764*x -0.033*x -0.226*x +0.689*x +0.474*x +0.309*x +2.747

Rule 2: IF AC is low [MF 2] AND HC is low [MF 2] AND RT is low [MF 2] AND OT is low [MF 2] AND SDI is low [MF 2] AND SEL is low [MF 2] AND RPT is high [MF 2] AND PCR is low [MF 2] AND PAM is low [MF 2] AND PHM is low [MF 2] AND OC is low [MF 2] THEN output = -0.029*x -0.058*x +0.006*x -0.059*x -0.104*x -0.001*x +0.006*x +0.050*x +0.002*x +0.009*x -0.009*x +0.054

Rule 3: IF AC is low [MF 3] AND HC is low [MF 3] AND RT is low [MF 3] AND OT is low [MF 3] AND SDI is low [MF 3] AND SEL is low [MF 3] AND RPT is high [MF 3] AND PCR is low [MF 3] AND PAM is low [MF 3] AND PHM is low [MF 3] AND OC is low [MF 3] THEN output = -4.253*x -1.033*x -0.507*x +8.283*x -0.153*x +0.201*x -0.056*x +0.462*x +0.033*x -0.106*x +0.088*x -0.425

The Takagi-Sugeno-Kang (TSK) type fuzzy rules obtained in the study reflect the system's behavioral tendencies towards a specific student profile. The antecedents of the three rules examined are identical: they describe a group of students with low experience and low past achievement, but whose current performance is on the rise.

The linear regression coefficients in the consequences of the rules determine the weight of the variables on the outcome.

- Rule 1 (High-Precision Estimation): This rule has the highest coefficient ($w_1 = -24.596$). Academically, this indicates that the model attempts to capture "extreme values" in this specific student group. It is the component of the system that gives the harshest punitive response to the increase in the number of erroneous trials.
- Rule 2 (Stable Estimation): The proximity of the coefficients to zero ($(w_i \approx 0)$) proves that this rule acts as an "anchor" in the system. It stabilizes the output by filtering noise in the variables and prevents overfitting.
- Rule 3 (Balanced Effect): Rule 3 establishes a balance between positive and negative weights. Specifically, the +8.283 coefficient on the fourth variable indicates that a particular parameter (e.g., dwell time or response speed) is considered strong evidence in favor of "learning" under this rule.

When the rules are examined as a whole, it is seen that the factors "Low Experience" and "Low Past Success" are treated as negative starting points (biases) by the model. However, the "RPT is high" condition causes the model to categorize the student as a "recovering student".

The fact that these three rules share the same antecedent confirms that the system uses a fuzzy clustering method and that the final prediction is produced by taking the weighted average of these rules (defuzzification). This shows that the model can make more flexible predictions by combining multiple local models with different gradients rather than a single linear model. The model's membership functions are determined by the model and shown in Fig. 2.

The membership functions used in the system are generally Gaussian or Generalized Bell-type curves. This choice ensures that the transitions between input values are smooth and continuous, rather than sharp, allowing the model to more accurately capture non-linear student behavior. As shown in Fig. 2, features such as HC and PAM show asymmetric intensity distributions at low values, whereas cognitive load indicators such as RT and SDI exhibit normal distributions. This configuration allows the system to respond aggressively to student errors while calculating success trends with a more balanced weighted average.

In the AC, HC, PAM, and PHM graphs, the curves are leaning to the left (towards zero). The model considers the amount of student use of systemic assistance tools as an indicator of "independence." The rapid damping of the curves after a scale of 0.4 indicates that the system labels a small number of trials as "positive/low" and that, when this threshold is exceeded, the student quickly moves into the "struggling student" category. The fact that the OC graph also peaks at similarly low values indicates that the initial responses to newly encountered tasks are weighted.

In the PCR and SEL graphs, membership levels span a wider range. These variables determine the model's "long-term student profile." In the PCR graph, high membership values in the 0.6-0.8 range indicate that the system evaluates students with moderate-to-high achievement levels within a similar confidence interval. In the SEL graph, the distinct separation of different clusters (blue, orange, and yellow curves) suggests that the model triggers different rule sets for novice and expert students.

The RT and SDI graphs exhibit a complete bell-shaped curve. These parameters represent the student's cognitive state. The RT graph peaking around 0.7-0.8 indicates that the system considers responses that follow a specific thought process "ideal," rather than those that are too fast (predictive) or too slow (disconnected). The SDI, which peaks in the 0.3-0.4 range, indicates that the system measures performance more stably on medium-easy tasks.

The RPT function is perfectly symmetrical and centered at 0. These variables measure the model's "learning momentum." Centering the momentum at zero indicates that the model has an unbiased starting point for the student's direction (progress or regression), and even a slight deviation in the trend immediately affects the output (via Rule 1-3 coefficients).

During the fuzzification of the 11 model input parameters, the variances of the Gaussian membership functions were optimized based on the educational significance of each variable.

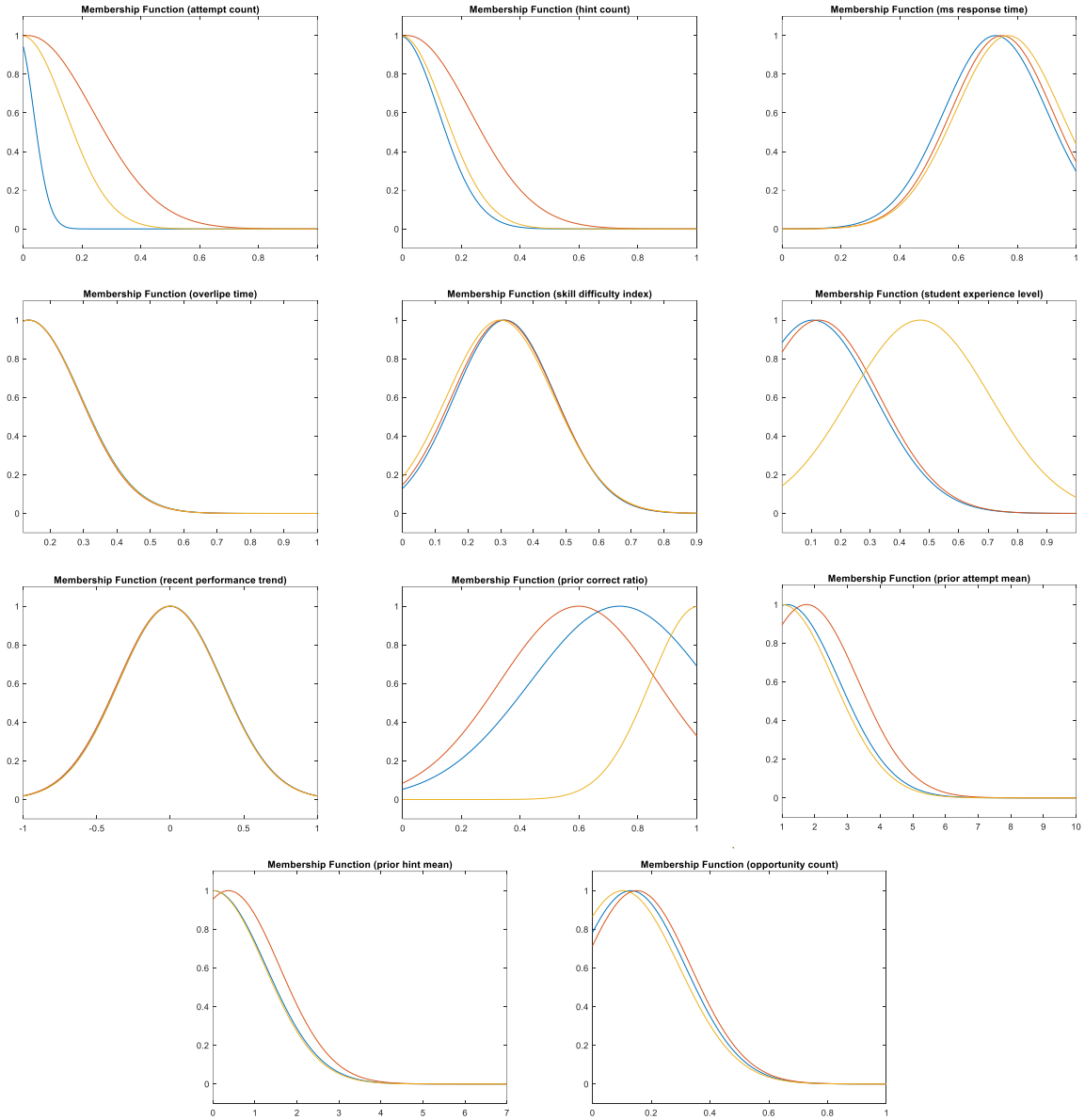


Fig. 2. Membership functions created with ANFIS and hybrid features ("AC", "HC", "RT", "OT", "SDI", "SEL", "RPT", "PCR", "PAM", "PHM", and "OC", respectively).

V. Discussion of Results

The findings of this study reveal, from both methodological and pedagogical perspectives, the advantages that hybrid approaches offer in predicting student achievement in online mathematics education compared to singular models and traditional assessment methods.

A. Hybrid Model Performance and Predictive Stability

Experimental results show that hybrid structures combining RF, XGBoost, and ANFIS models exhibit more stable performance in predicting continuous success scores (regression) and identifying at-risk students (classification). Specifically, the RF+Hybrid model, achieving the lowest RMSE (0.2369) and the highest R^2 (0.7327), confirms the power of tree-based ensemble learning algorithms in modeling complex, nonlinear student behavior. However, the weighted hybrid model, which achieves the highest F1-score (0.9413) and recall value (0.9544) in the classification phase, demonstrates that combining the models with weights minimizes erroneous decisions and produces more reliable early warning signals.

The numerical similarity in the classification performance of the models necessitates considering not only 'total accuracy' but also strategic metrics such as 'prediction stability' and 'error type cost' when selecting a model. While single-domain models (e.g., ANFIS-Raw) achieve high accuracy in some scenarios, the weighted hybrid model's high recall (0.9544) and F1 score (0.9413) during the classification phase provide a more reliable decision boundary for identifying at-risk students. This finding confirms that the hybrid architecture minimizes asymmetric error costs (mistaking failing students for passing students),

generating a more reliable early warning signal in the context of educational analytics.

B. The Role of Feature Engineering and Information Complementarity

Feature analysis results indicate that combining raw data and derived features is critical for model performance. The dramatic decrease in R^2 values (≈0.17–0.20) in scenarios where derived features are used alone indicates that these features cannot replace raw interaction information and should be included in the model as "context providers". In particular, the negligible correlation of the RPT feature with other features indicates that it provides independent, high-value "learning momentum" information for the model.

C. Uncertainty Management and Pedagogical Interpretability

The IF-THEN rules and membership functions provided by the ANFIS model have made the logical process underlying predictions transparent, unlike those of machine learning models, which are often described as "black boxes."

- Help-Seeking Behavior: The concentration of membership functions for HC and AC features at low values indicates that the model positively weights the use of a small number of hints as "learning independence" and quickly moves the learner to the "risky" category when a certain threshold is exceeded.
- Cognitive Load: The bell-curve form of RT documents that the system considers responses that come after a reasonable thought process as "ideal," rather than responses that are too fast (predictive) or too slow (disconnected).

D. Practical Applicability and Decision Support Potential

The study's two-stage (regression and classification) structure provides educators with both a continuous "probability of success" and a clear "successful/at-risk" label. The use of ROC-based adaptive thresholding created a more reliable decision boundary, particularly by reducing the pedagogical risks of false negatives (counting a failing student as successful). This suggests that the proposed model can be integrated into an early warning system for real-time monitoring on online mathematics platforms.

VI. Conclusion and Future Work

This study presents a unique hybrid framework combining fuzzy logic and machine learning techniques to predict student performance in online mathematics education. The main findings of the research are as follows:

- Superiority of the Hybrid Structure: Compared to single models, combining ANFIS's uncertainty modeling with the high predictive power of Random Forest and XGBoost yielded more balanced, reliable results.
- Prediction Accuracy: While the Random Forest (RF)-based hybrid model yielded the lowest regression error (RMSE: 0.2369), the Weighted Hybrid model excelled in classification accuracy (Accuracy: 0.9167) and risk detection sensitivity (Recall: 0.9544).
- Feature Interaction: Supporting raw interaction data with derived pedagogical features significantly increased the model's explanatory power (R^2).
- Explainability: The linguistic rules generated by ANFIS, unlike "black box" models, have enabled understanding of the cognitive and behavioral reasons behind the risk of failure (such as overuse of hints or temporal disconnections).

Research results have shown that different modeling paradigms (neuro-fuzzy and tree-based ensemble learning) achieve similar accuracy levels. However, the high predictive power offered by Random Forest and XGBoost models, supported by ANFIS's pedagogically interpretable IF-THEN rules, distinguishes the proposed hybrid framework from 'black box' models, making it unique. This not only provides high predictive performance but also enhances the pedagogical value of the study by enabling an understanding of the cognitive reasons behind the risk of failure (such as overuse of hints or temporal disconnections).

In the future, this model is planned to be tested with datasets from different disciplines and integrated into a real-time educational interface to provide teachers with instant intervention reports.

VII. Pedagogical Recommendations

Based on the findings from the model's membership functions and rule analysis (ANFIS), the following recommendations have been developed for online mathematics platforms:

- Balanced Use of Hints: Since excessive hinting (HC) is perceived by the model as a direct risk signal, the system should direct students to additional explanatory videos after a certain point instead of overwhelming them with hints.
- Encouragement of Ideal Thinking Time: Since very fast responses are often considered guessing (gaming the system), a minimum "waiting time" should be encouraged for students to read and analyze the question.
- Cumulative Follow-up: Due to the cumulative nature of mathematics education, the system should automatically define micro-tasks to "remind" students of past topics for those with low past achievement rates (PCR).
- Feedback Focused on Learning Momentum: Motivational messages should be delivered focusing not only on the result but also on the student's rising performance trend (RPT).

VIII. Limitations and Threats to Validity

The following limitations should be considered when evaluating the results of the study:

- Dataset Coverage: Analyses were performed only for the mathematics domain using the ASSISTments dataset (2009-2010); additional validations are needed to generalize the results to other subject areas (language, science, etc.).

- External Factors: Off-platform data, such as students' off-platform study habits, physical environment, or socio-economic status, could not be included in the model.
- Interaction Limitations: To maintain the stability of the model, students with very few (below 5) or many (above 1000) interactions were excluded from the analysis, as this may affect performance in extreme profiles.
- Computational Cost: The hybrid structure combining ANFIS and ensemble models may require high computational power in systems with very large-scale and real-time applications.

REFERENCES

[1] Z. Shou, M. Xie, J. Mo, and H. Zhang, "Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach," *Appl. Sci.*, vol. 14, no. 6, p. 2522, 2024.

[2] J. Stojanović *et al.*, "Application of distance learning in mathematics through adaptive neuro-fuzzy learning method," *Comput. Electr. Eng.*, vol. 93, p. 107270, 2021.

[3] W. Zhang, X. Huang, S. Wang, J. Shu, H. Liu, and H. Chen, "Student performance prediction via online learning behavior analytics," in *2017 International Symposium on Educational Technology (ISET)*, IEEE, 2017, pp. 153–157. Accessed: Jan. 24, 2026.

[4] A. Prasanth, "TabNet–XGBoost Hybrid Model for Student Performance Prediction and Customized Feedback.," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 10, 2025.

[5] E. Evangelista, "A hybrid machine learning framework for predicting students' performance in virtual learning environment," *Int. J. Emerg. Technol. Learn. IJET*, vol. 16, no. 24, pp. 255–272, 2021.

[6] J. Wang and Y. Yu, "Machine learning approach to student performance prediction of online learning," *PloS One*, vol. 20, no. 1, p. e0299018, 2025.

[7] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model. User-Adapt. Interact.*, vol. 19, no. 3, pp. 243–266, 2009.

[8] X. Zhang, Y. Zhang, A. L. Chen, M. Yu, and L. Zhang, "Optimizing multi label student performance prediction with GNN-TINet: A contextual multidimensional deep learning framework," *PloS One*, vol. 20, no. 1, p. e0314823, 2025.

[9] Z. R. Khudhair, M. Raheema, and J. Salman, "Evaluation of the Use of ANFIS in Predicting Student's," *Kerbala J. Eng. Sci.*, vol. 3, no. 3, pp. 101–113, Sep. 2023, doi: 10.63463/kjes1088.

[10] F. Widyahastuti and V. U. Tjhin, "Student Performance Prediction using Online Behavior Discussion Forum with Data Mining Techniques:," in *Proceedings of the Borneo International Conference on Education and Social Sciences*, Banjarmasin, Indonesia: SCITEPRESS - Science and Technology Publications, 2018, pp. 90–95. doi: 10.5220/0009017000900095.

[11] V. Ponnan, A. Abdulla, H. Muthukumaran, and V. Natarajan, "Data-Driven Hybrid Fuzzy-ML Model for Comprehensive Student Assessment," in *Understanding Uncertainty: Modern Approaches in Fuzzy System and Applications*, 2025, pp. 100–123.

[12] "Predicting Student Academic Success Using Deep Learning: A Multi-Factor Approach to Performance Prediction," *J. Logist. Inform. Serv. Sci.*, Mar. 2025, doi: 10.33168/JLISS.2025.0114.

[13] M. Nazir, A. Noraziah, M. Rahmah, M. Fakherldin, and A. Khawaji, "Transforming Education with Deep Learning: A Systematic Review on Predicting Student Performance and Critical Challenges," *Fusion Pract. Appl.*, no. Issue 2, pp. 79–99, Jan. 2025, doi: 10.54216/FPA.180207.

[14] A. E. Villegas-Espinoza and J. I. Necochea-Chamorro, "Using Deep Learning in Student Performance Prediction: A Systematic Review," *TEM J.*, vol. 14, no. 3, p. 2472, 2025.

[15] M. A. A. Bakar, A. T. Ab Ghani, M. L. Abdullah, N. Ismail, and S. Ab Aziz, "Adaptive Neuro-Fuzzy Inference System (ANFIS) Formulation to Predict Students' Neuroscience Mechanistic: A Concept of an Intelligent Model to Enhance Mathematics Learning Ability.," *TEM J.*, vol. 11, no. 4, 2022.

[16] I. Sapuguh, N. Ahlina, A. Wahyudi, B. Setyawan, and A. S. Rosalinda, "Development of fuzzy logic based student performance prediction system," *J. Tek. Inform. CIT Medicom*, vol. 15, no. 6, pp. 284–290, 2024.

[17] A. Al-Hmouz, J. Shen, R. Al-Hmouz, and J. Yan, "Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning," *IEEE Trans. Learn. Technol.*, vol. 5, no. 3, pp. 226–237, 2011.

[18] E. B. George, "Explainable AI Methods for Predicting Student Grades and Improving Academic Success," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 23s, pp. 117–126, Mar. 2025, doi: 10.52783/jisem.v10i23s.3680.

[19] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Explainable AI and Fuzzy Logic Systems," in *Theory and Practice of Natural Computing*, vol. 11324, D. Fagan, C. Martín-Vide, M. O'Neill, and M. A. Vega-Rodríguez, Eds., in Lecture Notes in Computer Science, vol. 11324. , Cham: Springer International Publishing, 2018, pp. 3–20. doi: 10.1007/978-3-030-04070-3_1.

[20] A. Altaher and O. BaRukab, "Prediction of student's academic performance based on adaptive neuro-fuzzy inference," *Int. J. Comput. Sci. Netw. Secur. IJCSNS*, vol. 17, no. 1, p. 165, 2017.

[21] T. O. Soyoye, T. Chen, and R. Hill, "Predicting Academic Performance of University Students Using Adaptive Neuro Fuzzy Inference System (ANFIS)- Subtractive Clustering Algorithm (ANFIS-SC): A Case Study in the UK," in *Advances in Computational Intelligence Systems*, vol. 1462, H. Zheng, D. Glass, M. Mulvenna, J. Liu, and H. Wang, Eds., in Advances in Intelligent Systems and Computing, vol. 1462. , Cham: Springer Nature Switzerland, 2024, pp. 315–333. doi: 10.1007/978-3-031-78857-4_24.

[22] D. Yang and M. S. A. Malik, "Design of performance evaluation method for higher education reform based on adaptive fuzzy algorithm," *PeerJ Comput. Sci.*, vol. 11, p. e3090, 2025.

[23] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, p. 237, 2020.

[24] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Sci. Rep.*, vol. 15, no. 1, p. 26879, 2025.

[25] C. S. Sanaboina and D. R. Sri, "Comparative Analysis of Hybrid Machine Learning Models for Predicting Student Performance with Data Balancing Via SMOTE and GAN," *Int. J. Eng. Res. Technol.*, vol. 14, no. 9, Sep. 2025, doi: 10.5281/zenodo.18096036.

[26] S. Sayadi and M. K. Sayadi, "A hybrid of random forest and SVM for predicting student performance," *Int. J. Ind. Eng. Manag. Sci.*, vol. 9, no. 1, pp. 44–51, 2022.

[27] C. Li and Z. Cao, "Deep learning-based AI model for predicting academic success and engagement among physical higher education students," *Sci. Rep.*, 2025.

[28] R. Hemidov, "Optimizing Student Performance Prediction Employing Hybrid Random Forest Models Boosted by Nature-Inspired Algorithms," *J. Artif. Intell. Syst. Model.*, vol. 3, no. 03, pp. 55–70, 2025.

[29] J. A. Goguen, "LA Zadeh. Fuzzy sets. Information and control," *J. Symb. Log.*, vol. 38, no. 4, pp. 656–657, 1973.

[30] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, May 2004, doi: 10.1080/08839510490442058.

[31] T. Bannet, S. Kumar, and K. Chidananda, "Machine Learning techniques for student performance prediction," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 10, pp. c616–c620, 2021.

[32] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Comput. Educ.*, vol. 61, pp. 133–145, 2013.

[33] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[34] O. Taylan and B. Karagözoğlu, "An adaptive neuro-fuzzy model for prediction of student's academic performance," *Comput. Ind. Eng.*, vol. 57, no. 3, pp. 732–741, 2009.

[35] V. O. Eguavoen and E. Nwelih, "A Hybrid FCM-PSO-ANFIS Model for Predicting Student Academic Performance," *J. Sarj. Tek. Inform.*, vol. 12, no. 3, pp. 91–98, 2024.

[36] K. O. Adefemi, M. B. Mutanga, and V. Jugoo, "Hybrid Deep Learning Models for Predicting Student Academic Performance," *Math. Comput. Appl.*, vol. 30, no. 3, p. 59, 2025.

[37] S. Göktepe Körpeoğlu, A. Filiz, and S. Göktepe Yıldız, "AI-driven predictions of mathematical problem-solving beliefs: Fuzzy logic, adaptive neuro-fuzzy inference systems, and artificial neural networks," *Appl. Sci.*, vol. 15, no. 2, p. 494, 2025.

[38] M. Zerkouk, M. Mihoubi, and B. Chikhaoui, "A Comprehensive Review of AI-based Intelligent Tutoring Systems: Applications and Challenges," Jul. 25, 2025, *arXiv*: arXiv:2507.18882. doi: 10.48550/arXiv.2507.18882.

[39] A. Kobsa, "User modeling and user-adapted interaction," in *Conference companion on Human factors in computing systems - CHI '94*, Boston, Massachusetts, United States: ACM Press, 1994, pp. 415–416. doi: 10.1145/259963.260532.

[40] N. T. Heffernan and C. L. Heffernan, "The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 470–497, Dec. 2014, doi: 10.1007/s40593-014-0024-x.

[41] M. Gray, "Predicting Student Performance Using Discussion Forums' Participation Data," Master's Thesis, Duke University, 2024. Accessed: Jan. 24, 2026.

[42] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[43] Z. A. Pardos and N. T. Heffernan, "Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing," in *User Modeling, Adaptation, and Personalization*, vol. 6075, P. De Bra, A. Kobsa, and D. Chin, Eds., in Lecture Notes in Computer Science, vol. 6075. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 255–266. doi: 10.1007/978-3-642-13470-8_24.

[44] R. S. Baker, T. Martin, and L. M. Rossi, "Educational Data Mining and Learning Analytics," in *The Handbook of Cognition and Assessment*, A. A. Rupp and J. P. Leighton, Eds., Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016, pp. 379–396. doi: 10.1002/9781118956588.ch16.

[45] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. Syst. Man Cybern.*, vol. 23, no. 3, pp. 665–685, 1993.

[46] J. Ferrandiz, D. Fonseca, and A. Banawi, "Mixed Method Assessment for BIM Implementation in the AEC Curriculum," in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, Eds., Cham: Springer International Publishing, 2016, pp. 213–222. doi: 10.1007/978-3-319-39483-1_20.

[47] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression," in *21st Annual SAS Malaysia Forum, 5th September*, Kuala Lumpur, 2007.

[48] D. Duan and S. Zhang, "Hybrid Model of Fuzzy Logic and Recurrent Neural Network for Dynamic Student Achievement Prediction," *Informatica*, vol. 49, no. 26, 2025.

[49] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data Cogn. Comput.*, vol. 8, no. 12, p. 187, 2024.

[50] W. J. Van Der Linden and R. K. Hambleton, "Item Response Theory: Brief History, Common Models, and Extensions," in *Handbook of Modern Item Response Theory*, W. J. Van Der Linden and R. K. Hambleton, Eds., New York, NY: Springer New York, 1997, pp. 1–28. doi: 10.1007/978-1-4757-2691-6_1.

[51] D. J. Hand, "Handbook of modern item response theory," *Biometrics*, vol. 54, no. 4, p. 1680, 1998.

[52] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[53] T. Chen, "XGBoost: A Scalable Tree Boosting System," *Cornell Univ.*, 2016.

[54] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[55] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013, doi: 10.1089/big.2013.1508.

[56] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950, doi: 10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3.

[57] R. Pelánek, "Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques," *User Model. User-Adapt. Interact.*, vol. 27, no. 3, pp. 313–350, 2017.

[58] X. Guo, R. Wang, X. Sun, and Q. Zhang, "Research on Student Performance Prediction Based on SVM Optimized by Hybrid SSA.," *IAENG Int. J. Comput. Sci.*, vol. 52, no. 7, 2025.

[59] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, vol. 1857, in Lecture Notes in Computer Science, vol. 1857. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[60] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2025.

[61] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[62] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[63] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996, doi: 10.1007/BF00117832.

An Empirical Analysis of Google Play Data Safety Disclosures: A Consistency Study of Privacy Indicators in Mobile Gaming Apps

Bakheet Aljedaani
Computer Science Department-Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—The Google Play marketplace has introduced the Data Safety section to improve transparency regarding how mobile applications (apps) collect, share, and protect user data. This mechanism requires developers to disclose privacy and security-related practices, including data collection, data sharing, and data protection measures. However, the reliability of these disclosures depends on developer self-reporting, raising concerns about their accuracy. This study investigates the consistency between developer-reported Data Safety disclosures and observable privacy indicators extracted from Android application packages (APKs). An empirical analysis was conducted on a dataset of 41 mobile gaming apps, including 21 children-oriented and 20 general-audience apps. A static analysis approach was used to extract key privacy indicators, including device identifiers, data sharing practices, personal information access, and location access. These indicators were systematically compared with corresponding disclosures using a structured consistency evaluation framework. The results reveal varying levels of agreement across privacy categories. Device identifier disclosures show relatively high consistency (87.8%), whereas other indicators exhibit substantial mismatches. In particular, location-related disclosures show the highest inconsistency rate (56.1%), followed by personal information and data sharing indicators. Comparative analysis shows similar mismatch patterns across app categories. Chi-square tests further indicate that these differences are not statistically significant, suggesting that inconsistencies are not associated with app category but reflect broader challenges within the analyzed mobile gaming dataset. These findings highlight limitations in current marketplace transparency mechanisms and emphasize the need for improved validation approaches to ensure accurate privacy reporting.**

Keywords*—Android security; data safety; mobile gaming apps; privacy analysis; static analysis

I. Introduction

Mobile applications (apps) have become an integral part of modern digital ecosystems, supporting a wide range of services including healthcare, entertainment, education, and commerce. To enable these services, mobile apps frequently collect and process user data for purposes (e.g., analytics, advertising, personalization, and service improvement) [1], [2]. As a result, concerns regarding the transparency of data collection and sharing practices have become increasingly important. End-users often rely on information presented in app marketplaces to understand how their data may be handled before installing an app. However, determining whether such disclosures accurately reflect actual app behavior remains a challenging problem [3]. Prior research has shown that mobile apps often access sensitive resources and user data in ways that are not always clearly communicated to end-users, raising concerns about privacy awareness and informed consent [4], [5]. In addition, studies have highlighted that permission systems alone do not provide sufficient insight into how user data is actually processed or shared within apps [6], [7].

To address these concerns, Google introduced the Data Safety section in the Google Play marketplace, requiring developers to disclose how their apps collect, share, and protect user data [8]. This feature provides structured information about data practices, including whether apps collect sensitive data (e.g., location or personal information, whether data is shared with third parties, and whether it is encrypted during transmission). The introduction of Data Safety labels represents an important step toward improving transparency and enabling end-users to make more informed decisions when selecting mobile apps. However, a recent study suggests that privacy labels may not fully address users' actual privacy concerns and found that a significant portion of users' privacy questions remain unanswered or only partially addressed by existing mobile app privacy labels [9]. Similar transparency initiatives have been explored in prior work, including efforts to improve privacy notices, standardized disclosure formats, and machine-readable privacy representations [1], [10], and [11]. However, the effectiveness of such mechanisms depends on the accuracy and completeness of developer-reported disclosures, which may vary depending on developer interpretation and implementation of reporting requirements [12]. Despite these improvements, the reliability of developer-reported privacy disclosures remains uncertain. This challenge is further compounded by the difficulty developers face in accurately completing privacy labels, particularly when apps integrate complex third-party components and data processing mechanisms [13]. The information presented in Data Safety labels is self-reported and not automatically verified by the platform, raising concerns about potential inconsistencies between declared practices and actual app behavior. Recent work has demonstrated that it may not fully reflect observable app behavior and has found that a substantial number of mobile apps transmit data that is not declared in their privacy labels, highlighting the lack of validation in current marketplace transparency mechanisms [14]. Previous studies have demonstrated discrepancies between privacy policies and observable app characteristics, including permission usage,

data flows, and third-party tracking behaviors [1], [15]. For example, studies have shown that many mobile apps request permissions that are not clearly justified by their functionality or fail to disclose certain data practices in their privacy policies [16], [17]. Other work has highlighted the role of third-party libraries and advertising frameworks in introducing additional data collection capabilities that may not be fully understood or disclosed by developers [15], [18], and [19]. These findings suggest that transparency mechanisms based solely on developer self-reporting may not always provide an accurate representation of app privacy behavior.

In this context, this study investigates the consistency between developer-reported Google Play Data Safety disclosures and observable privacy indicators extracted from Android Application Packages (APKs). Unlike prior work that focuses primarily on permissions or privacy policies, this study provides a direct comparison between marketplace disclosures and application-level technical indicators, enabling a more comprehensive assessment of privacy transparency. The analysis focuses on mobile gaming apps, one of the most widely adopted categories of mobile apps, particularly among children and young end-users [20], [21]. Given the scale and popularity of mobile games, as well as their increasing reliance on analytics and advertising infrastructures, understanding the reliability of their privacy disclosures is of significant importance. To provide deeper insights, the study includes both children-oriented and general-audience mobile gaming apps, enabling a comparative analysis of disclosure practices across app categories. In addition, the study applies statistical analysis techniques, including Chi-square tests, to evaluate whether observed differences in disclosure mismatches are statistically significant. By combining static analysis of APK files with structured analysis of marketplace disclosures, this study aims to provide empirical evidence regarding the effectiveness of Data Safety labels as a transparency mechanism in mobile app ecosystems. This study addresses the following research questions (RQs):

RQ1: To what extent do Google Play Data Safety disclosures align with observable static privacy indicators in mobile gaming apps?

This RQ aims to measure the level of consistency between developer-reported privacy disclosures and indicators extracted through static analysis of APK files.

RQ2: Do disclosure inconsistencies differ between children-oriented and general-audience apps?

This RQ aims to examine whether app category influences the likelihood of mismatches between static indicators and Data Safety labels.

RQ3: Are the observed differences in disclosure mismatches statistically significant?

This RQ aims to apply statistical analysis techniques to determine whether disclosure inconsistencies are associated with app category or occur broadly across the dataset.

To answer these RQs, we conducted an empirical study combining static analysis of Android apps with structured analysis of Google Play Data Safety disclosures. The study was performed on a selected dataset of 41 widely installed Android mobile gaming apps, including both children-oriented and general-audience titles. The dataset was divided into two groups: 21 children-oriented apps and 20 general-audience apps, selected based on popularity indicators such as installation counts and marketplace rankings to ensure representation of widely used apps with large user bases. The study was carried out through a four-phase workflow, including two main technical parts: static analysis and Data Safety extraction. The static analysis was performed for APK files to extract privacy-related indicators, including device identifier usage, personal information access, data sharing signals, and location access, based on permissions and observable app artifacts. Then, we manually collected and analyzed the corresponding Google Play Data Safety disclosures, mapping developer-reported information into structured indicators aligned with the static analysis results. The extracted data were then systematically compared to identify mismatches between observable indicators and reported disclosures, followed by comparative analysis across app categories and statistical validation using Chi-square testing. The results reveal that while certain indicators, such as device identifiers, exhibit relatively high agreement, other categories, including location, personal information, and data sharing, show substantial inconsistencies between developer-reported disclosures and observable app characteristics. Furthermore, similar mismatch patterns were observed across both children-oriented and general apps, suggesting that privacy transparency challenges are not limited to specific app categories but may reflect broader issues in marketplace disclosure mechanisms. We believe this study makes the following key contributions:

1) Empirical comparison of privacy disclosures and technical indicators: This work provides a systematic comparison between developer-reported Google Play Data Safety disclosures and observable static privacy indicators, enabling a direct assessment of the reliability of marketplace transparency mechanisms.

2) Cross-category analysis of mobile gaming apps: The study analyzes both children-oriented and general-audience apps, providing insights into whether privacy transparency differs across app categories subject to different regulatory and design considerations.

3) Identification of systematic disclosure inconsistencies: The results reveal consistent mismatch patterns across multiple privacy indicators, highlighting potential limitations in current self-reported disclosure mechanisms within mobile application marketplaces.

4) Integration of statistical analysis for validation: The study applies Chi-square statistical testing to evaluate whether observed differences are statistically significant, providing a rigorous empirical basis for interpreting the findings.

5) Implications for improving privacy transparency mechanisms: The findings provide actionable insights for developers, platform providers, and researchers, emphasizing the need for improved tooling, clearer reporting guidelines, and potential automated validation mechanisms.

The rest of this study is organized as follows: Section II presents the most relevant works to our study. Section III details the adopted research methodology. Section IV reports the study findings, followed by a discussion in Section V. Threats to validity are discussed in Section VI, and Section VII concludes the study and suggests potential future work.

II. Related Work

Research on mobile app privacy and security has been widely explored from multiple perspectives, including permission analysis, data flow analysis, privacy policies, third-party tracking, and transparency mechanisms. This section reviews relevant studies and positions the current work within the existing literature.

A. Permission-Based Privacy Analysis in Mobile Apps

Early research on mobile privacy primarily focused on analyzing Android permission systems as indicators of app behavior. Felt et al. in [4] conducted one of the earliest systematic studies of Android permissions, highlighting the gap between permission requests and user understanding. Similarly, Kelley et al. in [5] investigated end-user perception of permission requests, showing that end-users often struggle to interpret their implications. Subsequent studies demonstrated that permissions alone are insufficient to accurately characterize privacy risks, as apps may request unnecessary permissions or use them in unexpected ways [22], [7]. Further work extended permission-based analysis by examining how permissions relate to app functionality and privacy exposure. Zhou and Jiang in [17] analyzed Android malware and demonstrated how permissions can be abused for malicious purposes. Enck et al. in [23] introduced static analysis techniques to identify potential privacy leaks in Android apps. However, several studies have concluded that permission-based approaches do not represent privacy behavior with sufficient precision, as they do not capture how data is processed or shared after being accessed [24], [25].

B. Static and Dynamic Analysis of Privacy Behavior

To overcome the limitations of permission-based approaches, researchers have developed static and dynamic analysis techniques to examine how apps handle end users' data. Tools such as FlowDroid [16] and TaintDroid [23] enabled precise tracking of sensitive data flows within Android apps. These techniques have been widely used to detect privacy leaks and unauthorized data transmissions. Dynamic analysis approaches, such as those proposed by Rastogi et al. in [26] and Ren et al. in [27] have further demonstrated that many apps transmit sensitive data to remote servers, often without clear end-user awareness. While these approaches provide deeper insights into app behavior, they are often computationally expensive and difficult to scale across large datasets. In contrast, static analysis remains a widely used technique due to its scalability and ability to analyze large numbers of apps efficiently [28]. However, both approaches face challenges in capturing behaviors introduced by obfuscation, dynamic code loading, and third-party libraries.

C. Third-Party Libraries and Tracking Ecosystems

A significant body of research has examined the role of third-party libraries, particularly advertising and analytics SDKs, in shaping privacy risks in mobile apps. Studies by Grace et al. in [29] and Wang et al. in [19] showed that third-party components frequently introduce additional data collection and tracking capabilities that are not always transparent to users. Similarly, Ikram et al. in [30] analyzed mobile advertising ecosystems and found widespread tracking behavior across apps. Recent studies have highlighted that developers may not fully understand the privacy implications of integrated SDKs, leading to incomplete or inaccurate disclosures of data practices [31], [32]. These findings are particularly relevant in the context of mobile games, where monetization strategies often rely heavily on advertising and analytics infrastructures. As a result, third-party libraries represent a key factor contributing to the complexity of privacy transparency in mobile ecosystems.

D. Privacy Policies and Transparency Mechanisms

Another line of research has focused on evaluating privacy policies and user-facing transparency mechanisms. Zimmeck et al. in [10] developed automated techniques for analyzing privacy policies and identifying inconsistencies between stated and actual practices. Cranor in [11] discussed the challenges of designing effective privacy notices, emphasizing that traditional policy formats are often difficult for users to understand. More work has explored structured and machine-readable approaches to privacy disclosures, aiming to improve usability and transparency [33], [34]. However, even when structured disclosures are available, users may still struggle to interpret their meaning or assess their implications [35]. These challenges highlight the importance of evaluating not only the presence of privacy disclosures but also their accuracy and reliability.

E. Studies on Google Play Data Safety Labels

With the introduction of Google Play Data Safety labels, research has begun to examine their effectiveness as a transparency mechanism. While this feature is relatively recent, emerging studies suggest that privacy labels may face important limitations in both accuracy and usability. For instance, Koch et al. in [14] conducted a large-scale empirical study analyzing the consistency between Apple's privacy labels and actual app behavior through network traffic analysis. Their results showed that a significant number of apps transmit data not declared in their privacy labels, indicating discrepancies between developer-reported disclosures and observable behavior. The study also highlighted that privacy labels are not systematically validated during the app review process, raising concerns about their reliability as a transparency mechanism. Zhang and Sadeh in [9] evaluated the effectiveness of mobile app privacy labels in addressing users' privacy questions through an empirical analysis of a large corpus of user concerns. Their findings indicate that existing privacy labels often fail to provide sufficient information to answer many user questions, with a substantial portion of questions either unanswered or only partially addressed. These results suggest that, beyond issues of accuracy, privacy labels may also suffer from limitations in completeness and usability.

Cross-platform analyses have also highlighted inconsistencies in privacy label disclosures across different mobile ecosystems. Khandelwal et al. in [36] compared

privacy labels of apps available on both Android and iOS platforms and found notable differences in how data practices are reported across marketplaces. Their findings suggest that privacy disclosures may vary depending on platform-specific guidelines and developer interpretations, further raising concerns about the consistency and reliability of privacy labels. Lin et al. in [12] conducted a user-centered study comparing Android Data Safety labels with iOS privacy labels and found that, although some end-users perceive the labels as helpful, many users find them vague, incomplete, and difficult to interpret, leading to misunderstandings of app data practices. The study also reported that end-users often miss critical information due to interface design and express uncertainty regarding the reliability of developer-provided disclosures.

Furthermore, Zimmeck et al. in [10] and Reidenberg et al. in [35] demonstrated that privacy disclosures and policies often fail to accurately reflect actual data practices, highlighting limitations in user-facing transparency frameworks. Similarly, empirical analyses of mobile apps have shown that developers may over-report or under-report data collection practices, partly due to ambiguity in reporting requirements and limited visibility into third-party library behavior [19], [37]. In addition, prior research has emphasized the significant role of third-party analytics and advertising libraries in shaping app privacy behavior. Studies such as Han et al. in [19] and Derr et al. in [32] showed that embedded SDKs can introduce additional data collection mechanisms that are not always clearly reflected in developer disclosures. As a result, privacy labels may capture potential data practices associated with integrated components rather than directly observable behavior within the app package. These findings suggest that, while Data Safety labels represent a promising step toward improving transparency, they may still face challenges similar to those identified in earlier studies on privacy policies and mobile app disclosures.

Positioning of our Study: While prior work has extensively examined mobile privacy from perspectives such as permissions, data flows, third-party tracking, and privacy policies, relatively few studies have focused on the direct comparison between marketplace transparency mechanisms and observable application-level indicators. Existing studies on Data Safety labels remain limited and often do not integrate systematic static analysis with structured evaluation of disclosure consistency. This study addresses this gap by providing a configuration-aware and indicator-based comparison between developer-reported Data Safety disclosures and static privacy indicators extracted from APK files. In addition, it extends prior work by analyzing both children-oriented and general-audience apps, and by applying statistical analysis techniques to evaluate the significance of observed differences. By combining insights from software analysis and marketplace transparency research, this work contributes a more comprehensive understanding of the reliability of privacy disclosures in modern mobile ecosystems.

III. Research Methodology

This study adopts an application-centric static analysis approach to examine the consistency between Google Play Data Safety disclosures and observable technical indicators within Android apps. The methodology follows a structured empirical workflow consisting of four main phases: 1) design the study protocol, 2) APK collection and static inspection, 3) Data Safety label collection from Google Play, and 4) data analysis. Fig. 1 presents an overview of the research workflow, illustrating all the steps undertaken to conduct this study.

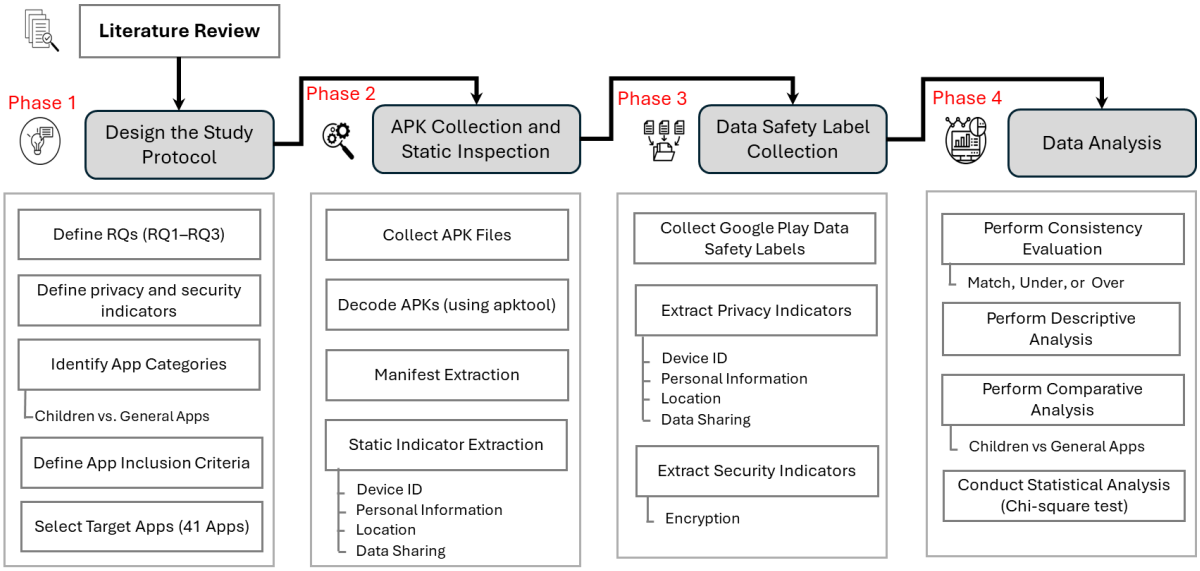


Fig. 1. An overview of the research workflow.

A. Phase 1: Designing the Study Protocol

The first phase focuses on establishing a systematic protocol to guide data collection and analysis. The protocol was designed based on prior literature on mobile privacy analysis and platform transparency mechanisms. First, the RQs (RQ1–RQ3), presented in Section I, were formulated to define the scope and objectives of the study. Based on these questions, a set of observable privacy and security indicators

was defined to guide the extraction and comparison process. This phase also involved the following:

1) Privacy and security indicators: To systematically evaluate disclosure accuracy, the study defined a set of observable indicators derived from Android application artifacts and Data Safety disclosures. These indicators were informed by prior research on Android privacy and security analysis, particularly studies focusing on permission usage and data access behaviors [38], [39], [40], and [41]. The selected indicators include:

- Location data collection.
- Device identifiers (Device ID).
- Personal information access.
- Data sharing practices.
- Encryption in transit (as declared in the Data Safety label).

These indicators were selected to enable a systematic comparison between static app characteristics and Google Play Data Safety disclosures. By focusing on these categories, the study provides a structured way to evaluate whether developer-reported privacy practices align with observable indicators within the app binaries, offering a more comprehensive perspective on privacy transparency beyond permission-based analysis alone.

2) Identifying app categories: To support comparative analysis, the selected apps were categorized into two groups: children-oriented apps and general-audience mobile games. This categorization was based on Google Play age ratings, store descriptions, and developer-provided metadata. The classification enables the investigation of potential differences in disclosure practices across app types.

3) App inclusion criteria: To enable a meaningful empirical evaluation, applications were selected based on a set of predefined inclusion criteria. Specifically, the selected apps must: 1) belong to the mobile gaming category in the Google Play marketplace, 2) demonstrate high popularity based on installation counts, user ratings, and store rankings, ensuring that the dataset reflects widely used applications, 3) be clearly classified as either children-oriented or general-audience, and 4) have an accessible APK file that can be successfully analyzed using static analysis techniques.

4) App selection: Based on these criteria, a dataset of 41 Android mobile gaming apps was constructed, consisting of 21 children-oriented apps and 20 general-audience games. The selected apps represent a diverse set of genres and development frameworks within the mobile gaming ecosystem. To support responsible reporting and avoid potential bias, applications are referenced using anonymized identifiers (i.e., *App1-App41*), with the mapping between identifiers and actual app names provided in the Appendix.

B. Phase 2: APKs Collection and Static Inspection

The second phase involved four main steps: 1) APKs collection, 2) APKs decoding, 3) manifest extraction, and 4) static indicator extraction. This phase involved the following:

1) APKs collection: APK files for the selected apps were obtained from publicly available sources to enable static analysis. In some cases, apps were distributed in split formats (e.g., XAPK or APKM), requiring additional processing to reconstruct usable APK files. Several downloaded packages were incomplete or corrupted, as indicated by identical file sizes or extraction errors. These issues were addressed by re-downloading the apps from alternative sources while ensuring consistency with the selected app versions. This process ensured that all analyzed APKs were valid and suitable for further analysis. The static analysis process employed multiple Android reverse-engineering and inspection tools. APK decoding and resource extraction were conducted using apktool version 2.9.3, while permission and manifest inspection were supported using Android Asset Packaging Tool (AAPT) from Android SDK Build-Tools version 36.1.0. Additional validation and package inspection were performed using JADX version 1.5.5 to examine application resources and package structures. These tools were selected due to their widespread adoption in Android software analysis research and their suitability for large-scale static analysis tasks.

2) APKs decoding: The collected APK files were decoded using static analysis tools (i.e., apktool) to enable access to their internal resources and configuration files. The decoding process extracted readable representations of app components, including the manifest file and resource directories. In some cases, decoding failures occurred due to obfuscation or packaging inconsistencies, requiring repeated attempts or the use of alternative tool configurations. Successfully decoded APKs were then used as the basis for extracting privacy-related attributes.

3) Manifest extraction: The AndroidManifest.xml file was extracted from each decoded APK as the primary source of application-level configuration and permission declarations. This file was systematically inspected to identify privacy-relevant attributes, including declared permissions, app components, and network-related configurations. Variations in manifest structure across apps occasionally required manual verification to ensure accurate interpretation. The extracted manifest data provided a consistent foundation for identifying observable privacy indicators across all apps.

4) Static indicator extraction: Static privacy indicators were derived from the extracted manifest data by mapping specific permissions and configuration attributes to predefined categories, including device IDs, personal information, location access, and data sharing. The extraction process involved both automated parsing and manual validation to ensure accuracy and consistency across apps.

C. Phase 3: Collection of Google Play Data Safety Labels

This phase involved collecting Data Safety labels for each app, which are publicly available on the Google Play marketplace. The Google Play Data Safety section was manually inspected to record developer-declared privacy practices. Based on the study protocol, the following disclosure fields were extracted: 1) location data collection, 2) device identifier collection, 3) personal information collection, 4) app activity tracking, and 5) encryption in transit. All extracted information was systematically recorded in a structured dataset to ensure consistency and enable reliable comparison across apps. The complete dataset is provided in the Appendix.

D. Phase 4: Data Synthesis And Comparative Analysis

The final phase of the methodology focuses on analyzing the extracted data to evaluate the consistency between static indicators and Data Safety disclosures. This phase includes 1) consistency evaluation, 2) descriptive analysis, 3) comparative analysis, and 4) statistical testing to identify patterns and assess the significance of observed discrepancies.

1) Consistency evaluation: All extracted attributes were organized into a structured dataset consisting of 41 app records and multiple privacy indicators. For each indicator, a systematic comparison was performed between the Google Play Data Safety disclosures and the corresponding static analysis evidence derived from APK inspection. To characterize the relationship between reported and observable practices, three comparison outcomes were defined as follows:

- Match: The static analysis evidence is consistent with the developer-reported Data Safety disclosure.
- Under-disclosure: The static analysis indicates the presence of data access or collection that is not declared in the Data Safety label.
- Over-disclosure: The Data Safety label declares data collection or usage that is not supported by observable static evidence.

This step forms the core of the analysis by establishing a consistent framework for identifying discrepancies between developer disclosures and observable application characteristics.

2) Descriptive statistics: Descriptive statistics were used to quantify the prevalence of the defined comparison outcomes across the dataset. The analysis focused on overall mismatch frequency, indicator-specific discrepancy rates, and distribution patterns across privacy categories. This step enables the identification of systematic inconsistencies between marketplace disclosures and observable application characteristics, providing an initial understanding of how disclosure accuracy varies across different types of data.

3) Comparative analysis: To further investigate potential differences across app types, the dataset was divided into two categories: children-oriented apps and general-audience mobile games. Comparative analysis was then performed to examine variation in mismatch rates across these categories for each privacy indicator. This step supports the evaluation of whether disclosure inconsistencies differ depending on app audience, directly addressing the study's comparative research objectives.

4) Statistical analysis: To assess whether the observed differences between app categories are statistically significant, Chi-square tests of independence were performed using Microsoft Excel [42]. The test evaluates the relationship between categorical variables, specifically disclosure outcome (match vs. mismatch) and app category (children-oriented vs. general-audience). By applying this statistical method, the study moves beyond descriptive observations and provides a rigorous assessment of whether observed differences reflect meaningful patterns or occur due to random variation. This step strengthens the validity of the findings and supports more robust conclusions regarding the nature of disclosure inconsistencies.

5) Ethical considerations: The study relies exclusively on static inspection of publicly distributed Android APKs. No runtime interaction with user accounts or personal data was performed. All results are reported in aggregate form, and apps are anonymized to minimize potential harm to developers or users.

IV. Findings

This section reports the observed security and privacy indicators across the analyzed mobile gaming apps. The results are presented to highlight patterns in data collection, data sharing, and configuration-related practices extracted through static analysis. In addition, the findings examine the level of consistency between these observable indicators and the corresponding Data Safety disclosures. This analysis provides an empirical basis for identifying potential mismatches and understanding their distribution across different privacy categories.

A. Consistency Between Static Indicators and Data Safety Labels

The comparison between static indicators and Data Safety disclosures revealed varying levels of consistency across privacy categories. As shown in Table I, the highest level of consistency was observed for device ID disclosures, with an agreement rate of 87.8% and a mismatch rate of 12.2% (i.e., 7.3% for under-disclosure, 4.9% for over-disclosure). In contrast, lower levels of consistency were observed for the remaining indicators. Data sharing practices exhibited a mismatch rate of 36.6%, while personal information disclosures showed a higher mismatch rate of 46.3% (i.e., 29.3 for under-disclosure and 17.0% for over-disclosure). The largest discrepancy was observed for location-related disclosures, where 56.1% of apps showed inconsistencies between static analysis results and the corresponding Data Safety labels. These findings indicate that the accuracy of privacy disclosures varies considerably depending on the type of data involved. The relatively high consistency observed for device identifiers may be attributed to the direct relationship between advertising identifiers and well-defined platform permissions, making them easier for developers to report accurately.

TABLE I. SUMMARY OF DISCLOSURE CONSISTENCY ACROSS PRIVACY INDICATORS

Indicator	Match	Under-disclosure	Over-disclosure
Device ID	87.8%	7.3%	4.9%
Personal Info	53.7%	29.3%	17.0%
Location	43.9%	0%	56.1%
Data Sharing	63.4%	36.6%	0%

TABLE II. EXAMPLE COMPARISON BETWEEN STATIC INDICATORS AND DATA SAFETY LABELS (FIRST 10 APPS)

App ID	Device ID (S/L)	Personal Info (S/L)	Location (S/L)	Data Sharing (S/L)
App1	Y / N	Y / N	N / N	Y / N
App2	Y / Y	Y / Y	N / N	Y / N
App3	Y / Y	Y / Y	N / Y	Y / Y
App4	Y / Y	Y / N	N / N	Y / N
App5	Y / Y	Y / N	N / N	Y / N
App6	Y / Y	Y / N	N / Y	Y / N
App7	Y / Y	Y / N	N / Y	Y / Y
App8	Y / Y	Y / Y	N / Y	Y / N
App9	Y / N	Y / N	N / N	Y / N
App10	Y / Y	Y / Y	N / N	Y / N

S denotes static analysis results, while L denotes developer-reported Data Safety disclosures; Y=Yes, N=No.

In contrast, categories such as location, personal information, and data sharing often involve more complex interactions with third-party libraries and backend services, which may introduce ambiguity in how data practices are interpreted and disclosed. As a result, these categories are more prone to inconsistencies between developer-reported information and observable app characteristics. To provide a more concrete illustration of these discrepancies, Table II presents a detailed comparison between static indicators and Data Safety labels for a subset of the analyzed apps (first 10 apps). This example demonstrates how mismatches occur at the individual app level, including cases of both under-disclosure and over-disclosure across different privacy indicators.

B. Comparison Between Children and General Games

To better understand whether disclosure inconsistencies differ across application types, the dataset was divided into two categories: children-oriented apps and general-audience mobile games. The mismatch rates for each privacy indicator across the two categories are summarized in Table III. The results indicate that the relationship between app category and disclosure inconsistency varies across privacy indicators rather than following a uniform pattern. As shown in Table III, the mismatch rate for device identifier disclosures was 19.0% for children-oriented apps, compared to 5.0% for general-audience apps. Similarly, data sharing disclosures showed mismatch rates of 42.9% for children-oriented apps and 30.0% for general-audience apps. In contrast, personal information disclosures exhibited a slightly higher mismatch rate among general applications (50.0%) compared to children-oriented apps (42.9%). A similar trend was observed for location-related disclosures, where general applications showed a higher mismatch rate (60.0%) than children-oriented apps (52.4%). These results indicate that while some indicators (e.g., device identifiers and data sharing) show higher mismatch rates in children-oriented applications, others (e.g., personal information and location) are more inconsistent in general applications. Overall, disclosure inconsistencies are observed across both categories, suggesting that such discrepancies are not confined to a specific type of application but rather reflect broader challenges in accurately reporting privacy practices within the Data Safety framework.

TABLE III. COMPARISON OF MISMATCH RATES BETWEEN APP CATEGORIES

Indicator	Children App Mismatch	General App Mismatch
Device ID	19.0%	5.0%
Personal Info	42.9%	50.0%
Location	52.4%	60.0%
Data Sharing	42.9%	30.0%

C. Statistical Significance Analysis

To determine whether the observed differences in disclosure mismatches between children-oriented and general applications are statistically significant, a Chi-square test of independence was performed for each privacy indicator. For each indicator, contingency tables were constructed comparing the number of matching and mismatching cases across the two app categories. The Chi-square test was selected because it is suitable for examining relationships between categorical variables, such as app category and disclosure consistency outcomes. In addition, the test enables the identification of potential associations between mismatch patterns and app categories while providing a statistically grounded interpretation of the observed differences. The results of the statistical analysis are summarized in Table IV. As shown in Table IV, none of the examined indicators exhibits a statistically significant association between app category and disclosure mismatch. Specifically, device identifier mismatches yielded $\chi^2 = 0.804$ ($p = 0.370$), data sharing mismatches yielded $\chi^2 = 0.281$ ($p = 0.596$), personal information mismatches yielded $\chi^2 = 0.021$ ($p = 0.885$), and location mismatches yielded $\chi^2 = 0.031$ ($p = 0.860$). These findings suggest that disclosure inconsistencies are not dependent on application category, but rather reflect broader limitations in the accuracy of marketplace privacy disclosures within the analyzed mobile gaming apps.

TABLE IV. CHI-SQUARE TEST RESULTS FOR DISCLOSURE CONSISTENCY

Indicator	χ^2 Value	p-value
Device ID	0.804	0.370
Personal Info	0.021	0.885
Location	0.031	0.860
Data Sharing	0.281	0.596

D. Security Practices in Data Safety Labels

In addition to privacy-related disclosures, the analysis also examined security-related information reported in the Data Safety labels, particularly whether user data is encrypted during transmission. The results indicate variability in how mobile gaming apps report their data transmission security practices. While a majority of apps declared that data is encrypted in transit, a small subset of applications explicitly reported that user data is not encrypted in transit, including App35 and App40, as in Fig. 2. Although encryption is not directly observable through static analysis, this finding provides additional insight into how developers communicate security practices through the Data Safety framework. Encryption of data in transit is widely considered a fundamental security measure for protecting user data from interception and unauthorized access [43], [44]. The observed variability suggests that, similar to privacy disclosures, security-related declarations may also lack consistency, highlighting an additional dimension of transparency challenges within mobile app marketplaces.

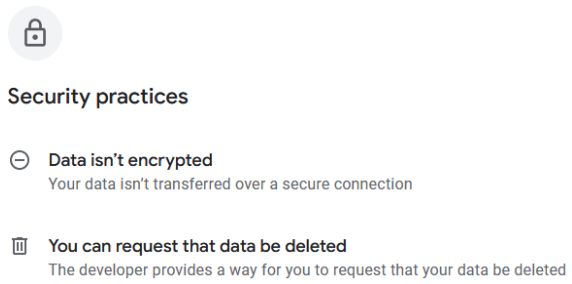


Fig. 2. Examples of apps reporting that data is not encrypted in transit in Google Play Data Safety labels.

V. Discussion

This section discusses the key findings of the study. It is important to note that the analysis is based on observable static indicators extracted from application artifacts rather than runtime behavior. Accordingly, the interpretations reflect potential privacy exposures inferred from static analysis rather than dynamic data flows.

A. Reliability of Data Safety Labels

The results of this study indicate that the reliability of Google Play Data Safety disclosures varies across different types of privacy indicators. Device identifier disclosures demonstrated the highest level of consistency between developer-reported labels and observable static indicators, with an agreement rate of 87.8%. This relatively high agreement suggests that developers may find it easier to accurately report the use of device identifiers, as these practices are typically associated with well-defined permissions and standardized advertising frameworks. In contrast, other indicators exhibited substantially higher mismatch rates, particularly for location (56.1%), personal information (46.3%), and data sharing (36.6%). A closer examination of mismatch types reveals distinct patterns across indicators. Location-related discrepancies were primarily driven by over-disclosure, where developers report data collection that is not supported by observable static evidence. Conversely, data sharing inconsistencies were predominantly associated with under-disclosure, indicating that data exchange mechanisms may not always be fully reported in Data Safety labels. Personal information disclosures exhibited a combination of both under- and over-reporting. These findings suggest that the reliability of Data Safety disclosures is not uniform but depends strongly on the type of data being reported, reflecting differences in how developers interpret and implement reporting requirements.

These findings are consistent with prior research examining transparency mechanisms in mobile app ecosystems. Several studies have reported inconsistencies between developer-reported privacy policies and observable app behavior, particularly in relation to permission usage and data flows [10], [35]. Similarly, recent work analyzing Google Play Data Safety disclosures has suggested that developers may either over-report or under-report certain data practices due to ambiguity in the reporting guidelines or uncertainty about third-party library behavior [12]. The results of this study extend these observations by providing empirical evidence within the context of mobile gaming apps and by systematically comparing Data Safety disclosures with static privacy indicators extracted directly from APK files. This comparison highlights that while Data Safety labels represent an important step toward improving transparency, their reliability may still depend heavily on how developers interpret and implement the reporting requirements.

B. Privacy Transparency in Children-Oriented Apps

An important objective of this study was to investigate whether privacy disclosure inconsistencies differ between children-oriented and general-audience apps. The results indicate that the relationship between app category and disclosure inconsistency is not uniform across privacy indicators. For certain indicators, such as device identifiers and data sharing, children-oriented apps exhibited higher mismatch rates. For example, device ID mismatches were observed in 19.0% of children-oriented apps compared to 5.0% of general-audience apps, while data sharing mismatches were 42.9% and 30.0%, respectively. In contrast, personal information and location-related disclosures showed higher mismatch rates among general-audience apps. Specifically, personal information mismatches were 50.0% for general apps compared to 42.9% for children-oriented apps, while location-related mismatches were 60.0% and 52.4%, respectively. These results indicate that inconsistencies in personal information and location disclosures are more pronounced in general-audience applications.

Despite these variations, statistical analysis confirmed that none of these differences are statistically significant, indicating that disclosure inconsistencies are present across both categories. These findings suggest that privacy transparency challenges are not limited to a specific group of apps but instead reflect broader issues in how developers interpret and report data practices within the Data Safety framework. This observation is consistent with prior studies showing that both children-oriented and general-audience apps may incorporate complex data collection mechanisms, including third-party analytics and advertising libraries, which can lead to inconsistencies between declared and actual practices [19], [45].

C. Systemic Nature of Disclosure Inconsistencies

The absence of statistically significant differences between children-oriented and general-audience apps suggests that inconsistencies between developer disclosures and observable indicators may represent a broader issue observed within the analyzed mobile gaming apps. Rather than being associated with a specific app category, these inconsistencies appear to arise from structural characteristics of modern mobile software development. This observation is consistent with recent empirical work [14], which shows that discrepancies between privacy disclosures and actual application behavior are widespread across mobile ecosystems. The study reports that numerous apps violate their declared privacy labels by transmitting undeclared data, indicating that such inconsistencies are systemic rather than isolated cases. This interpretation is further supported by prior work [13], which highlights that developers often face challenges in accurately completing privacy labels due to the complexity of modern mobile apps and limited visibility into third-party components. In particular, developers may struggle to understand and correctly report data collection behaviors, and tool-supported approaches based on static analysis have been proposed to improve label accuracy. In practice, many mobile apps rely heavily on third-party components such as advertising SDKs, analytics libraries, and cloud-based services. These components may introduce data collection capabilities that are not always clearly reflected in app permissions or static artifacts. Consequently, developers may report potential data practices associated with these components even when the corresponding behaviors are not directly observable within the application package. Taken together, these factors help explain the presence of indicator-specific mismatch patterns observed in this study, suggesting that disclosure inconsistencies are influenced not only by application category but also by the nature and complexity of the data being reported.

This interpretation aligns with previous research highlighting the complex role of third-party libraries in mobile privacy risks. Studies have shown that advertising and analytics frameworks frequently introduce additional data flows and tracking mechanisms that developers may not fully control or understand [32], [19]. As a result, accurately describing all potential data practices in marketplace disclosures can be challenging. Furthermore, platform reporting guidelines may encourage developers to disclose possible data usage scenarios rather than strictly observable behaviors, which may contribute to the mismatch patterns observed in this study. The findings therefore suggest that improving the accuracy of marketplace transparency mechanisms may require stronger tool support or automated validation approaches capable of identifying discrepancies between developer disclosures and observable app characteristics.

D. Security Implications of Data Safety Disclosures

In addition to privacy-related indicators, the analysis also examined security-related information reported in the Data Safety labels, particularly whether user data is encrypted during transmission. The results revealed variability in how mobile gaming apps report their encryption practices. While many apps indicated that user data is encrypted during transmission, several apps reported that user data may not be encrypted in transit. Encryption of data in transit is widely considered a fundamental security practice for protecting sensitive information from interception or unauthorized access. The presence of apps that report the absence of encryption mechanisms therefore raises potential concerns regarding the protection of user data in mobile gaming environments.

Previous studies examining security practices in mobile apps have similarly reported inconsistent adoption of secure communication mechanisms, particularly among apps that rely on third-party advertising or analytics services [44], [23]. In some cases, developers may depend on external services to manage data transmission security, which may lead to variations in how encryption practices are implemented or reported. The variability observed in this study suggests that Data Safety disclosures may provide useful signals regarding security practices but may not always present a complete picture of app security. From a user's perspective, inconsistent reporting of encryption practices may make it difficult to assess the level of protection provided by different apps. These findings highlight the importance of continued research into automated methods for evaluating both privacy and security transparency in mobile app marketplaces.

E. Implications for Mobile Privacy Transparency

The findings of this study have several implications for improving transparency mechanisms in mobile app marketplaces. First, the observed inconsistencies between static indicators and Data Safety disclosures suggest that developers may face challenges in accurately interpreting the reporting requirements introduced by Google Play. A recent study has shown that even when privacy labels are available, they may not fully address users' information needs. Furthermore, the work demonstrated that many user privacy questions remain unanswered by current label designs, highlighting limitations in their ability to support informed decision- making [9]. In particular, the higher mismatch rates observed for indicators such as location access, personal information, and data sharing highlight the difficulty of mapping app behavior and third-party library functionality to the categories defined in the Data Safety framework. Similar challenges have been identified in prior studies examining the relationship between app permissions, privacy policies, and developer disclosures [39], [6]. These results suggest that additional guidance or automated tools may be required to assist developers in generating accurate Data Safety reports. For example, developer tools that automatically analyze app artifacts and recommend disclosure categories could reduce ambiguity and improve the reliability of privacy reporting. In particular, the distinct mismatch patterns observed across different indicators (e.g., over-disclosure in location and under-disclosure in data sharing) suggest that reporting challenges may be closely tied to how specific data categories are defined and interpreted within the Data Safety framework.

The results also have implications for platform providers and researchers investigating mobile privacy transparency. For platform providers such as Google Play, the mismatch patterns identified in this study suggest that self-reported privacy disclosures alone may not fully capture app behavior. Integrating automated verification mechanisms into the

marketplace submission process could help identify inconsistencies between declared data practices and observable app characteristics. For example, automated checks based on static or dynamic analysis could flag potential mismatches during app submission and prompt developers to review their disclosures. From a research perspective, the findings highlight the importance of combining multiple analysis approaches, including static analysis, dynamic analysis, and marketplace metadata analysis, in order to better understand privacy transparency in mobile ecosystems. Future research may extend this work by developing automated techniques for identifying discrepancies between app behavior and marketplace disclosures on a larger scale.

VI. Threats to Validity

As with any empirical study involving software artifacts and marketplace disclosures, several factors may influence the interpretation of the results. Following established guidelines for empirical software engineering research, threats to validity are categorized into internal, construct, and external validity.

A. Internal Validity

Internal validity concerns whether the observed results may be influenced by methodological limitations during data collection and analysis. In this study, privacy indicators were extracted through static analysis of APK files, which may not capture behaviors triggered dynamically at runtime. For example, certain privacy-relevant actions may occur through remote services or dynamically loaded components that are not visible within the app package. Similar limitations of static analysis approaches have been reported in prior Android security studies [10], [11]. Another potential threat relates to APK availability and format variations. Some apps were initially obtained in formats such as XAPK or APKM, requiring additional extraction steps before analysis. In several cases, incomplete packages or identical file sizes indicated invalid downloads, which required replacing a small number of apps while maintaining the dataset balance. To mitigate these risks, permissions were manually verified when automated extraction produced unexpected results.

Another limitation relates to the exclusive reliance on static analysis techniques. Although static inspection enables scalable analysis of application artifacts and observable privacy indicators, it does not capture runtime behaviors, dynamically triggered data flows, or network transmissions that may occur during application execution. Consequently, the findings reflect observable static indicators rather than confirmed runtime data practices. Future work incorporating dynamic analysis techniques may provide a more comprehensive assessment of disclosure consistency and runtime privacy behavior.

B. Construct Validity

Construct validity concerns whether the selected indicators accurately represent the privacy practices being studied. In this work, privacy behavior was approximated using static indicators derived primarily from Android permissions and observable APK artifacts. While these indicators provide insights into potential data access capabilities, they may not fully capture all privacy-related behaviors within mobile apps. Some forms of data collection may occur through embedded third-party libraries, analytics frameworks, or advertising SDKs that do not require explicit permissions. As a result, certain privacy practices may remain undetected through permission-based analysis. Additionally, the Google Play Data Safety labels represent developer-reported disclosures rather than independently verified measurements of app behavior. Developers may declare potential data practices associated with integrated libraries even when those behaviors are not observable in the APK. Similar challenges in interpreting developer privacy disclosures have been reported in previous research on mobile privacy transparency [23], [27].

C. External Validity

External validity concerns the generalizability of the study findings beyond the analyzed dataset. This study analyzed 41 mobile gaming apps, including both children-oriented and general-audience games. Although this sample size is comparable to that of many empirical studies in mobile privacy research, it represents only a subset of the broader Android application ecosystem. In addition, the relatively small group sizes (21 children-oriented apps and 20 general-audience apps) may reduce the statistical power of the Chi-square analysis. Therefore, the absence of statistically significant differences between app categories should be interpreted cautiously, as non-significant results may reflect limited sample size rather than the complete absence of category-specific effects. Another limitation relates to the scope of the analyzed applications. The study focuses exclusively on mobile games, which may exhibit different privacy and monetization practices compared to other application categories such as social networking, productivity, or financial apps. Furthermore, the results reflect a snapshot of apps and Data Safety disclosures at the time of data collection. Since mobile applications and marketplace disclosures continuously evolve, future updates may lead to different observations. Despite these limitations, the dataset includes widely used apps with millions of downloads, providing a meaningful snapshot of privacy disclosure practices within the mobile gaming ecosystem [19]. Future studies involving larger datasets, additional app categories, and dynamic analysis techniques may further validate and extend the findings presented in this work.

VII. Conclusion and Future Work

Mobile apps increasingly rely on complex data collection, analytics, and advertising infrastructures to support functionality and monetization. To improve transparency regarding these practices, the Google Play marketplace introduced the Data Safety section, requiring developers to disclose how apps collect, share, and protect user data. Although this initiative represents an important step toward improving privacy transparency, the reliability of developer-reported disclosures remains largely dependent on self-reporting and may not always reflect observable app behavior. In this study, we conducted an empirical comparison between developer-reported Data Safety disclosures and static privacy indicators extracted from APKs to assess the consistency between marketplace transparency mechanisms and observable app characteristics. The analysis focused on a dataset of mobile gaming apps, including both children-oriented and general-

audience games, enabling a comparative evaluation of privacy disclosure practices across app categories.

The study was conducted through a structured empirical process that involved extracting privacy-relevant indicators from APK files through static analysis and comparing these indicators with the corresponding disclosures reported in the Google Play Data Safety section. By examining indicators related to device identifiers, personal information access, data sharing practices, and location access, this work provides a systematic assessment of how accurately developer disclosures reflect observable privacy-related artifacts within mobile apps. The key findings of this study can be summarized as follows:

- Device identifier disclosures exhibited the highest level of agreement between static indicators and Data Safety labels, suggesting that some data practices may be easier for developers to report accurately due to their clear association with specific platform permissions and advertising frameworks.
- Other privacy indicators showed substantial inconsistencies, particularly for personal information, data sharing, and location-related disclosures, indicating that developer-reported transparency mechanisms do not always align with observable app artifacts.
- Children-oriented and general-audience apps exhibited similar patterns of disclosure inconsistencies, and statistical analysis revealed no significant differences between these categories, suggesting that transparency challenges extend across the broader mobile gaming ecosystem.
- Security-related disclosures, particularly regarding encryption practices, varied across apps, highlighting additional concerns regarding the clarity and reliability of information presented to users through the Data Safety framework.

These findings suggest that while the Google Play Data Safety section represents an important advancement in improving transparency in mobile app marketplaces, its effectiveness may be limited by the reliance on developer self-reporting and the inherent complexity of modern mobile app architectures. In particular, the widespread use of third-party analytics, advertising, and attribution frameworks may make it difficult for developers to precisely map app behavior to the categories defined within the Data Safety reporting framework. As a result, developers may disclose potential data practices associated with integrated libraries even when those practices are not directly observable in the app package.

Based on the empirical evidence, this study highlights several practical implications. First, developers may benefit from improved tool support and automated analysis mechanisms that assist in generating accurate Data Safety disclosures based on observable app artifacts. Second, platform providers such as Google Play may consider integrating automated validation mechanisms into the app submission process to detect inconsistencies between declared privacy practices and observable app characteristics. Such mechanisms could improve the reliability of marketplace transparency features and provide users with more accurate information when evaluating mobile apps. We believe this study opens several avenues for future research as follows:

- The static analysis presented in this work can be extended with dynamic analysis techniques to observe runtime behavior, network communication patterns, and actual data flows. Combining static and dynamic analysis would enable researchers to evaluate whether the inconsistencies observed between Data Safety disclosures and static indicators correspond to real privacy risks during app execution.
- Future studies could conduct large-scale measurements across hundreds or thousands of apps to evaluate the reliability of Data Safety disclosures across different app categories beyond mobile games, such as social networking, productivity, or financial apps.
- Longitudinal studies could examine how developer disclosures evolve as apps are updated, monetization models change, and platform reporting requirements continue to evolve. Such studies would provide valuable insights into the stability and long-term effectiveness of marketplace transparency mechanisms.
- Building on prior human-centered research, future work could explore how users interpret and rely on Data Safety labels when making installation decisions, and whether these disclosures meaningfully influence user perceptions of app privacy and security.

The results of this study can benefit multiple stakeholders. Researchers may use the findings to further investigate transparency mechanisms in mobile ecosystems and develop automated methods for detecting inconsistencies between app behavior and marketplace disclosures. Developers and app publishers may leverage the insights to improve privacy engineering practices and ensure that marketplace disclosures accurately reflect app behavior. Ultimately, this work contributes to a growing body of research indicating that improving transparency in mobile ecosystems requires not only clearer disclosure frameworks but also stronger technical mechanisms for validating the information presented to users. In addition to the empirical findings presented in this study, this work contributes to the growing body of research investigating privacy transparency in mobile app ecosystems by providing a systematic comparison between developer-reported marketplace disclosures and observable app characteristics. While previous studies have examined privacy policies, permission usage, and third-party tracking infrastructures in mobile apps, fewer studies have focused specifically on the consistency between marketplace transparency mechanisms and technical indicators extracted from app artifacts. By combining static analysis of APK files with structured analysis of Google Play Data Safety disclosures, this study provides a novel perspective on the reliability of developer-reported privacy information. The insights generated through this comparison highlight the importance of strengthening transparency mechanisms in mobile marketplaces and demonstrate the value of combining

software analysis techniques with marketplace metadata analysis in future privacy and security research.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgment

OpenAI's GPT-4 was minimally used for proofreading and grammar refinement during manuscript preparation.

References

[1] S. Saeed, "Usable privacy and security in mobile applications: perception of mobile end users in Saudi Arabia," Big Data and Cognitive Computing, vol. 8, p. 162, 2024.

[2] A. Alkinoon, T. C. Dang, A. Alghuried, A. Alghamdi, S. Choi, M. Mohaisen, et al., "A comprehensive analysis of evolving permission usage in Android apps: trends, threats, and ecosystem insights," Journal of Cybersecurity and Privacy, vol. 5, p. 58, 2025.

[3] D. Rodriguez, J. M. Del Alamo, C. Fernández-Aller, and N. Sadeh, "Sharing is not always caring: Delving into personal data transfer compliance in android apps," IEEE Access, vol. 12, pp. 5256-5269, 2024.

[4] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in Proceedings of the 18th ACM conference on Computer and communications security, 2011, pp. 627-638.

[5] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall, "A conundrum of permissions: installing applications on an android smartphone," in International conference on financial cryptography and data security, 2012, pp. 68-79.

[6] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang, "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing," in Proceedings of the 2012 ACM conference on ubiquitous computing, 2012, pp. 501-510.

[7] Y. Agarwal and M. Hall, "ProtectMyPrivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing," in Proceeding of the 11th annual international conference on Mobile systems, applications, and services, 2013, pp. 97-110.

[8] N. Alomar, J. Reardon, A. Girish, N. Vallina-Rodriguez, and S. Egelman, "The effect of platform policies on app privacy compliance: A study of child-directed apps," Proceedings on Privacy Enhancing Technologies 2025, 2025.

[9] S. Zhang and N. Sadeh, "Do privacy labels answer users' privacy questions," in Workshop on Usable Security and Privacy, 2023.

[10] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, et al., "Automated Analysis of Privacy Requirements for Mobile Apps," in NDSS, 2017, pp. 1.4-2.3.

[11] L. F. Cranor, "Necessary but not sufficient: Standardized mechanisms for privacy notice and choice," J. on Telecomm. & High Tech. L., vol. 10, p. 273, 2012.

[12] Y. Lin, J. Juneja, E. Birrell, and L. F. Cranor, "Data safety vs. app privacy: Comparing the usability of android and ios privacy labels," arXiv preprint arXiv:2312.03918, 2023.

[13] J. Gardner, Y. Feng, K. Reiman, Z. Lin, A. Jain, and N. Sadeh, "Helping mobile application developers create accurate privacy labels," in 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2022, pp. 212-230.

[14] S. Koch, M. Wessels, B. Altpeter, M. Olvermann, and M. Johns, "Keeping privacy labels honest," Proceedings on Privacy Enhancing Technologies, 2022.

[15] M. Huasong Meng, C. Yan, Q. Zhang, Z. Wang, K. Wang, S. Gee Teo, et al., "Assessing Privacy Compliance of Android Third-Party SDKs," arXiv e-prints, p. arXiv: 2409.10411, 2024.

[16] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, et al., "Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps," ACM sigplan notices, vol. 49, pp. 259-269, 2014.

[17] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in 2012 IEEE symposium on security and privacy, 2012, pp. 95-109.

[18] M. Ikram and M. A. Kaafar, "A first look at mobile ad-blocking apps," in 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), 2017, pp. 1-8.

[19] S. Han, J. Jung, and D. Wetherall, "A study of third-party tracking by mobile apps in the wild," Univ. Washington, Tech. Rep. UW-CSE-12-03, vol. 1, 2012.

[20] J. Pimienta, J. Brandt, T. Bethe, R. Holz, A. Continella, L. Jibb, et al., "Mobile apps and children's privacy: A traffic analysis of data sharing practices among children's mobile iOS apps," Archives of disease in childhood, vol. 108, pp. 943-945, 2023.

[21] R. Carlsson, S. Rauti, S. Laato, T. Heino, and V. Leppänen, "Privacy in Popular Children's Mobile Applications: A Network Traffic Analysis," in MIPRO ICT and Electronics Convention, 2023, pp. 1213-1218.

[22] J. Lin, B. Liu, N. Sadeh, and J. I. Hong, "Modeling {Users'} mobile app privacy preferences: Restoring usability in a sea of permission settings," in 10th Symposium On Usable Privacy and Security (SOUPS 2014), 2014, pp. 199-212.

[23] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, et al., "Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones," ACM Transactions on Computer Systems (TOCS), vol. 32, pp. 1-29, 2014.

[24] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, "Pscout: analyzing the android permission specification," in Proceedings of the 2012 ACM conference on Computer and communications security, 2012, pp. 217-228.

[25] D. Barrera, H. G. Kayacik, P. C. Van Oorschot, and A. Somayaji, "A methodology for empirical analysis of permission-based security models and its application to android," in Proceedings of the 17th ACM conference on Computer and communications security, 2010, pp. 73-84.

[26] V. Rastogi, Y. Chen, and X. Jiang, "Catch me if you can: Evaluating android anti-malware against transformation attacks," IEEE Transactions on Information Forensics and Security, vol. 9, pp. 99-108, 2013.

[27] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes, "Recon: Revealing and controlling pii leaks in mobile network traffic," in Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, 2016, pp. 361-374.

[28] D. Octeau, P. McDaniel, S. Jha, A. Bartel, E. Bodden, J. Klein, et al., "Effective {Inter-Component} communication mapping in android: An essential step towards holistic security analysis," in 22nd USENIX Security Symposium (USENIX Security 13), 2013, pp. 543-558.

[29] M. C. Grace, W. Zhou, X. Jiang, and A.-R. Sadeghi, "Unsafe exposure analysis of mobile in-app advertisements," in Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks, 2012, pp. 101-112.

[30] M. Ikram, H. J. Asghar, M. A. Kaafar, B. Krishnamurthy, and A. Mahanti, "Towards seamless tracking-free web: Improved detection of trackers via one-class learning," arXiv preprint arXiv:1603.06289, 2016.

[31] T. Book, A. Pridgen, and D. S. Wallach, "Longitudinal analysis of android ad library permissions," arXiv preprint arXiv:1303.0857, 2013.

[32] E. Derr, S. Bugiel, S. Fahl, Y. Acar, and M. Backes, "Keep me updated: An empirical study of third-party library updatability on android," in Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, 2017, pp. 2187-2200.

[33] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor, "A design space for effective privacy notices," in Eleventh symposium on usable privacy and security (SOUPS 2015), 2015, pp. 1-17.

[34] P. Emami-Naeini, H. Dixon, Y. Agarwal, and L. F. Cranor, "Exploring how privacy and security factor into IoT device purchase behavior," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1-12.

[35] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, et al., "Disagreeable privacy policies: Mismatches between

meaning and users' understanding," Berkeley Tech. LJ, vol. 30, p. 39, 2015.

[36] R. Khandelwal, A. Nayak, P. Chung, and K. Fawaz, "Comparing privacy labels of applications in android and ios," in Proceedings of the 22nd Workshop on Privacy in the Electronic Society, 2023, pp. 61-73.

[37] T. Book and D. S. Wallach, "A case of collusion: A study of the interface between ad libraries and their apps," in Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices, 2013, pp. 79-86.

[38] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, et al., "Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem," in The 25th annual network and distributed system security symposium (NDSS 2018), 2018.

[39] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," in Proceedings of the eighth symposium on usable privacy and security, 2012, pp. 1-14.

[40] J. Reardon, Á. Feal, P. Wijesekera, A. E. B. On, N. Vallina-Rodriguez, and S. Egelman, "50 ways to leak your data: An exploration of apps' circumvention of the android permissions system," in 28th USENIX security symposium (USENIX security 19), 2019, pp. 603-620.

[41] "OWASP Mobile Top 10 available at https://owasp.org/www-project-mobile-top-10/," 2023.

[42] V. University, "Conducting a chi squared contingency test using Excel [online] available at https://researchguides.library.vanderbilt.edu/c.php?g=156859&p=1171766," 2024.

[43] M. Georgiev, S. Iyengar, S. Jana, R. Anubhai, D. Boneh, and V. Shmatikov, "The most dangerous code in the world: validating SSL certificates in non-browser software," in Proceedings of the 2012 ACM conference on Computer and communications security, 2012, pp. 38-49.

[44] S. Fahl, M. Harbach, T. Muders, L. Baumgärtner, B. Freisleben, and M. Smith, "Why Eve and Mallory love Android: An analysis of Android SSL (in) security," in Proceedings of the 2012 ACM conference on Computer and communications security, 2012, pp. 50-61.

[45] I. Reyes, P. Wijesekera, J. Reardon, A. Elazari Bar On, A. Razaghpanah, N. Vallina-Rodriguez, et al., ""Won't somebody think of the children?" examining COPPA compliance at scale," in The 18th Privacy Enhancing Technologies Symposium (PETS 2018), 2018.

APPENDIX

The list of mobile gaming apps used in our study with the full results for each app and the extended version of Table II can be found in the following link: https://shorturl.at/mS5tW

Improving Emotion Recognition Accuracy Using a Multimodal Model (Face and Voice Video) Based on a Convolutional Neural Network (CNN)

Karnadi[1], Ermatita[2]*, Abdiansah[3]
Doctor of Engineering Program, Universitas Sriwijaya Palembang, Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya Palembang, Indonesia[2, 3]

***Abstract*—Advancements in Artificial Intelligence (AI) technology have enabled the recognition of human emotions. Along with the development of deep learning and multimodal processing methods, emotion analysis can now be performed by utilizing multiple data sources simultaneously, such as facial expressions and speech signals. However, existing emotion recognition systems still face limitations in terms of accuracy. This study aims to develop and evaluate a more accurate emotion recognition system by implementing a Convolutional Neural Network (CNN)-based prediction model that integrates facial and audio data simultaneously. The study utilizes the CREMA-D dataset, which consists of visual data in the form of facial images and audio data containing variations of emotional expressions. The research process includes data preprocessing, feature extraction, and multimodal integration using an optimized Convolutional Neural Network (CNN) architecture. The evaluation results based on the F1-score indicate that the multimodal facial and audio data enable the model to recognize emotions effectively. Model performance was measured using accuracy, precision, recall, and F1-score metrics. Experimental results show that the angry (ANG) class achieved the best performance with an F1-score of 82%, while the fear (FEA) class demonstrated the lowest performance with an F1-score of only 58%. The results further indicate that the multimodal model achieved higher accuracy than unimodal models, significantly improving generalization capability on diverse testing data. This study demonstrates an overall emotion recognition accuracy improvement of 69% through the combination of facial and audio features. The analysis of combined facial and speech features on emotion classification performance shows that the proposed model achieves good overall performance, where the integration of image and audio modalities improves the correctness of facial expression predictions. Future research is expected to further improve accuracy by incorporating additional modalities beyond facial and audio data.**

***Keywords*—Emotion recognition; CNN; multimodal learning; CREMA-D; face; voice; accuracy; deep learning**

I. Introduction

An important factor in human communication is emotion, [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], which influence behavior, perception, and decision-making. Research on automatic human emotion recognition has advanced rapidly thanks to advances in artificial intelligence, particularly the development of systems capable of understanding and responding to users' emotional states in a more human-like manner. To recognize human emotions more accurately, research has utilized single-modal observations [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] of either voice or facial data. Previous researchers [29], [30], [31], [32], [33], used the VGG Face architecture to extract facial features and MFCC (Mel-Frequency Cepstral Coefficients) for voice features. This system can be used to help develop more user-friendly and responsive human-computer interfaces (HCI) [34]. A three-layer system was developed, consisting of FEAM (facial analysis) [35], [36], [37], VSAM (voice analysis) [38], [39], [40], [41], [42], [43], and MERM (emotion merging for the metaverse) [44], [45], [46]. This system is designed to enhance the realism and accuracy of emotional interactions in virtual environments such as the metaverse. This system can identify seven primary emotions in real time by utilizing technologies such as FCNN (Fully Convolutional Neural Network) and MCycleGAN (Modified CycleGAN) [47], [48], [49]. In simulation experiments, the results were comparable to human perceptual capabilities, focusing on the use of voice and facial recognition technologies for social inclusion [50], [51], [52], [53], [54], [55], [56]. This system uses OpenCV, a Raspberry Pi, and facial recognition techniques such as Haar cascades and LBPH, as well as voice control, to open automatic doors. Test results show that this system is highly effective in helping people with disabilities access their homes independently, with an accuracy rate of 99.63%. All three studies indicate that the integration of auditory and visual data holds great potential for helping to develop intelligent systems that are responsive to real human needs [57]. The uncertainty that arises when using a single modality is one of the main challenges in emotion recognition. Audio signals are affected by ambient noise and the quality of the recording device, while lighting, the camera angle, or even deliberate facial expressions can influence the recording. Multimodal methods leverage the strengths of each modality and mitigate their weaknesses through the process of information fusion; this method offers a solution. In this study, we used CREMA-D, a crowdsourced multimodal dataset. This dataset consists of video recordings of actors expressing six basic emotions: happiness, sadness, anger, fear, frustration, and rage. This dataset is ideal for multimodal emotion recognition research because it includes visual, audio, and audiovisual data. To extract spatiotemporal features of facial expressions from the video data, a CNN-LSTM architecture was used. Meanwhile, the audio data was converted into Mel-spectrograms for feature extraction using a Convolutional Neural Network (CNN).

*Corresponding author.

Multimodal emotion recognition (MER) has emerged as a promising approach in this context because it can integrate various sources of emotional information, such as facial expressions and voice intonation, into a single approach. Research findings indicate that this multimodal method outperforms single-modal methods in detecting emotions.

Given these opportunities and challenges, this study aims to develop a multimodal model based on a Convolutional Neural Network (CNN) capable of integrating visual and audio information to improve the accuracy of emotion recognition.

II. Materials and Methods

A. Multimodal

Multimodal machine learning (also known as multimodal learning) is a subfield of machine learning that aims to develop and train models capable of utilizing various types of data and learning to link or combine these modalities, intending to improve predictive performance [58]. Deep learning (DL), as a cutting-edge technology, has achieved remarkable breakthroughs in many computer vision tasks due to its impressive capabilities in data representation and reconstruction. Naturally, it has been successfully applied to the field of multimodal hospital data fusion, yielding significant improvements over traditional methods. This survey aims to provide a systematic overview of deep learning-based multimodal hospital data fusion. More specifically, some key insights on this topic are presented first [59]. According to researchers [60], [61], [62], six machine learning models employing two distinct feature extraction techniques are considered foundational models.

B. Convolutional Neural Network Algorithm

A convolutional neural network is a type of neural network designed to process grid-based data, such as two-dimensional images. The term "convolution" refers to a linear algebraic operation that multiplies a filter matrix by the image to be processed [63], [64], [65], [66]. Fig. 1 shows the architecture of the proposed CNN model for facial expression recognition.

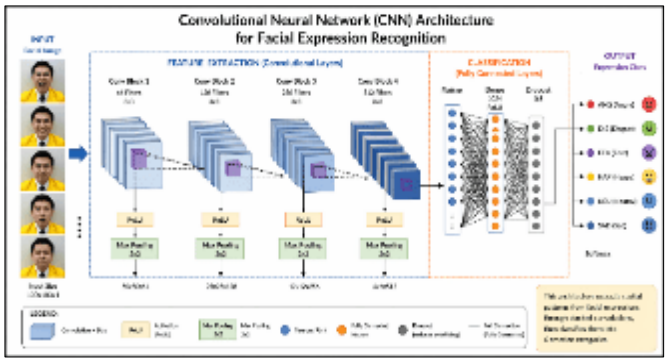


Fig. 1. Convolutional neural network architecture.

1) Data split (stratified split): Data splitting (stratified split) according to [66]. The Scikit-learn function used to split a dataset into two parts—training data and test data—is written in the Python code `train_test_split(...)`. In general, the splitting of a dataset into training and test data [67], [68], [69], [70], [71], can be expressed as follows:

$$D = \{(x_i, a_i, y_i)\}_{i=1}^{N} \quad (1)$$

$With: x_i, \epsilon\ X_{img}, a_i \in\ X_{audio}, y_i \in Y$

Divide D by:

$$D_{train} = \{(x_i, a_i, y_i)\}_{i\ =\ 1}^{0.8\ N}$$

$$D_{test} = \{(x_j, a_j, y_j)\}_{j\ =\ 1}^{0.8\ N}$$

$by\ stratification: P(y_i \in\ c_k)\ on\ D_{train} \approx P(y_i \in c_k)\ on\ D, \forall_{c_k} \in\ class$

2) Cutting to fit the batch: This function is used to reduce the amount of data by a multiple of the batch size. This is particularly important to avoid having an incomplete final batch, especially in models with multimodal inputs, such as images and audio recordings [72], [73], [74], [75].

The general formula is as follows:

$$n_{batch} = \left\lfloor \frac{N}{B} \right\rfloor \quad (2)$$

$$N' = n_{batch} . B$$

$$D' = \left\{ \left(x_i^{(1)}, x_i^{(2)}, y_i\right) \right\}_{i\ =\ 1}^{N'}$$

3) Training model: This formula is used to evaluate model performance during both training and testing [75], [76], [77]. The loss function plays a crucial role in the backpropagation process for updating the weights of a neural network through optimization (such as gradient descent) [78], [79]. It implicitly uses that average formula behind the scenes. Here is the formula used:

$$Loss = \frac{1}{N} \sum_{i=1}^{N} L(\hat{y}i, yi) \quad (3)$$

4) Predictions and evaluations: Predictions and evaluations are performed using the final-class prediction formula in a classification model that employs the softmax activation function. This formula is the standard formula used in multiclass classification with deep neural networks, particularly in the output layer of classification models that use the softmax activation function [80], [81], [82]. Here is the formula used:

$$y^{\wedge} = \text{argmax}(\text{softmax}(z)) \quad (4)$$

Proposed approach

C. Research Methods and Types

The objective of this experimental quantitative study, which employs a deep learning-based systems engineering approach, is to develop and evaluate a multimodal emotion recognition system using Convolutional Neural Networks (CNN).

D. Data Sources and Data Types

This study uses secondary data in the form of the CREMA-D dataset (Crowd-sourced Emotional Multimodal Actors Dataset), which is a publicly available multimodal dataset

accessible via the Kaggle platform (https://www.kaggle.com/datasets/orvile/crema-d-emotional-multimodal-dataset). The CREMA-D dataset consists of 7,442 video clips featuring 91 actors (48 men and 43 women) aged 20–74 who express six basic emotion categories: anger, happiness, sadness, fear, disgust, and neutral. Each sample in the dataset has been manually labeled and contains two main modalities: visual data in the form of facial expressions recorded in video format, and audio data in the form of voice intonation that can be extracted into acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC). This dataset was selected for its high annotation quality and good inter-rater reliability, its representative demographic diversity of actors, and its multimodal format, which aligns with the research objectives of developing a CNN model that integrates visual and audio information for emotion regulation identification.

E. Research Phases

The research process consists of several stages carried out by the author, including: data collection, data preprocessing, multimodal feature fusion, model training, model evaluation, and system implementation, as shown in Fig. 2 below [83], [84].

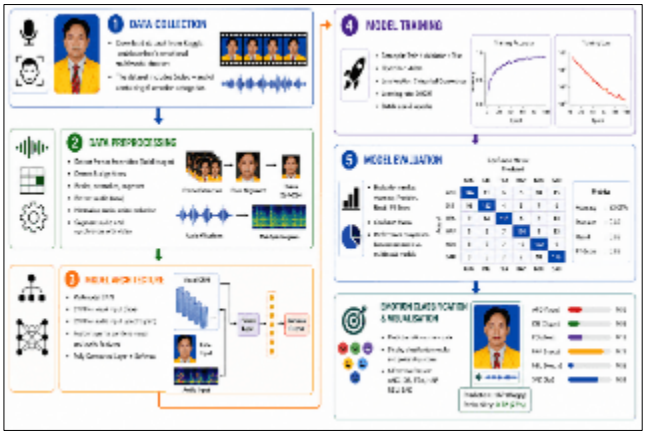


Fig. 2. Research phases.

The data collection phase involved downloading the CREMA-D dataset from the Kaggle platform, which contains 7,442 multimodal video clips across six emotion categories (anger, happiness, sadness, fear, disgust, neutral). This dataset was selected because it provides two modalities—facial expressions (visual) and voice intonation (audio)—both of which are well-labeled. The preprocessing stage involves extracting data from video files into two separate components. For the visual modality, face detection, alignment, cropping, and resizing of facial images to 64×64 pixels are performed. For the audio modality, MFCC (Mel-Frequency Cepstral Coefficients) features are extracted and converted into a spectrogram with dimensions of 40×128. Data augmentation is used to increase the diversity of the dataset and prevent overfitting.

This stage designs a dual-stream-based multimodal CNN architecture consisting of an Image Branch for processing facial images and an Audio Branch for processing audio spectrograms. Both branches use three convolutional blocks (32, 64, 128 filters) with Batch Normalization and Max Pooling. Features from both modalities are combined using concatenation in the Fusion Layer, then processed through fully connected layers for final classification.

The model was trained using a dataset split into a stratified split (70:15:15 for training, validation, and testing). The Adam optimizer with a learning rate of 0.001, a categorical cross-entropy loss function, and early stopping was used to optimize the network parameters. Training was conducted for up to 100 epochs with a batch size of 32 until the model reached convergence.

Model performance was evaluated using test data and various metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. A comparison was conducted between the multimodal model and unimodal models (visual-only and audio-only) to assess the effectiveness of the fusion strategy. Statistical analysis was performed to validate the significance of the performance improvement.

The final stage presents visualizations of the classification results, including a confusion matrix, graphs showing training and validation loss and accuracy, and examples of predictions on the test set. The visualizations also include Grad-CAM to interpret the facial regions that most influence the predictions and feature importance analysis to understand the contribution of each modality to the model's classification decisions.

F. CNN Model Architecture

The proposed architecture employs a dual-stream Convolutional Neural Network (CNN) approach with two separate branches to independently process visual and audio modalities. The visual branch is designed to extract facial expression features, while the audio branch utilizes a 40×128 MFCC spectrogram as the representation of speech features. The extracted feature representations from both modalities are subsequently fused through a concatenation mechanism at the fusion layer to construct an integrated multimodal representation before the emotion classification process. Furthermore, the dataset was divided using a stratified split strategy with a ratio of 70:15:15 for training, validation, and testing sets, respectively, in order to preserve class distribution balance and ensure a reproducible experimental setup. The emotion classification process using these results is shown in Fig. 3 below [85], [86], [87], [88], [89], [90], [91], [92], [93], [94].

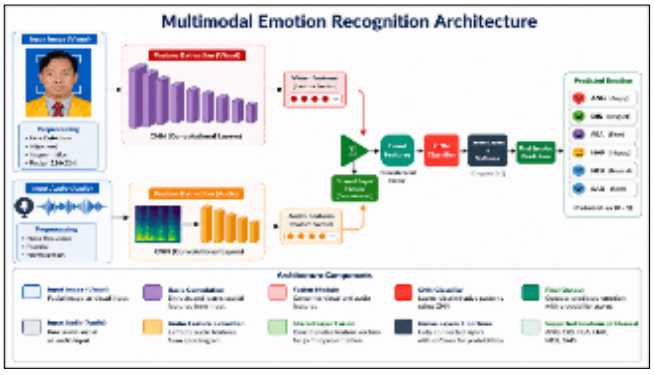


Fig. 3. CNN architecture.

Fig. 3 above shows the model architecture, which consists of two parallel branches for processing visual and audio modalities independently. In the Image Branch, 64×64×3 image inputs are

processed through three convolutional blocks with progressively larger filters (32, 64, 128), each equipped with Batch Normalization and Max Pooling. Global Average Pooling then aggregates the feature maps, followed by a 128-neuron Dense layer with Dropout, producing a 128-dimensional image feature vector.

In the Audio Branch, a 40×128×1 audio spectrogram is processed using a CNN architecture with the same filter configuration (32, 64, 128). Following Global Average Pooling, a dense layer with 128 neurons and Dropout produces a 128-dimensional audio feature vector that encodes the prosodic and spectral characteristics of the audio.

The Fusion Layer combines the two feature vectors via concatenation, producing a 256-dimensional multimodal representation. This representation is processed through two Dense layers (with 256 and 128 neurons) using Batch Normalization and Dropout to learn complex interactions between modalities. The Output Layer uses a Dense layer with a Softmax activation function to generate a probability distribution for each emotion regulation class, enabling the model to learn optimal representations and fusion strategies end-to-end.

TABLE I. DUAL STREAM CNN ARCHITECTURE FOR MULTIMODAL EMOTION RECOGNITION

Layer	Output Shape	Parameter
Input Image	255	238
Conv2D + ReLU	255	271
MaxPooling2D	255	145
Conv2D + ReLU	255	175
MaxPooling2D	215	229
MaxPooling2D	(32,32,64)	Pool size 2×2
GlobalAveragePooling	(64)	-
Input Audio MFCC	(40,128,1)	MFCC spectrogram
Conv2D + ReLU	(40,128,32)	Kernel 3×3
MaxPooling2D	(20,64,32)	Pool size 2×2
Conv2D + ReLU	(20,64,64)	Kernel 3×3
MaxPooling2D	(10,32,64)	Pool size 2×2
GlobalAveragePooling	(64)	-
Concatenate	(128)	Feature fusion
Dense + ReLU	(128)	Fully connected
Dropout	(128)	Rate = 0.5
Output Layer	(6)	Softmax activation

Table I presents the proposed dual-stream CNN architecture for multimodal emotion recognition, which consists of separate visual and audio processing branches. The visual branch processes RGB facial images through multiple convolutional and max-pooling layers to extract facial expression features, while the audio branch processes MFCC spectrogram inputs using a similar CNN structure to extract acoustic features. The extracted features from both modalities are then combined using a concatenation layer for multimodal feature fusion. The fused representation is further processed through fully connected and dropout layers before being classified into six emotion classes using a Softmax output layer. This architecture is designed to effectively learn complementary information from facial expressions and speech signals to improve emotion recognition performance.

G. Hardware and Software

Programming Language Tools and Software: Python Framework: TensorFlow/Keras or PyTorch Hardware: Nitro ANV16-71 Laptop with Intel(R) Core(TM) i7-14650HX processor, GPU, and 16 GB of RAM Supporting Software: OpenCV, Librosa, and Scikit-learn.

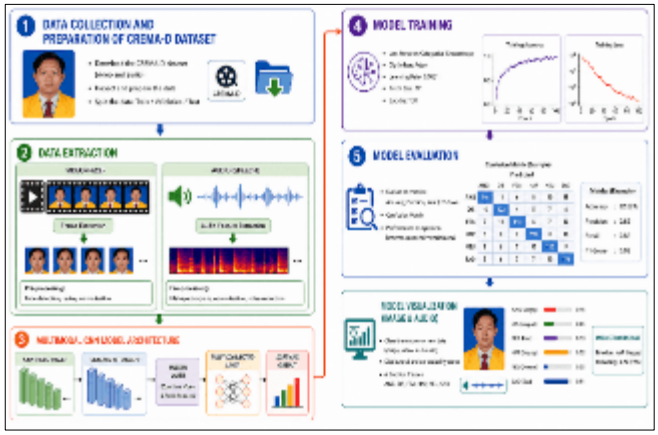


Fig. 4. Flowchart of the stages of multimodal deep learning analysis using the CNN algorithm.

Fig. 4 above illustrates the workflow of the multimodal emotion recognition system. The process begins with data collection, which includes video and audio. Both types of data then proceed to the preprocessing stage: video preprocessing to process facial frames and audio preprocessing to process audio signals. After that, the video and audio are processed by the CNN Feature Extract module to extract key features from each modality. The resulting features are then combined in the feature fusion stage, allowing the model to obtain a more comprehensive representation of emotions. Finally, the combined results are classified in the emotion classification stage to produce emotion categories.

III. RESULTS AND DISCUSSION

A. Data Collection

The dataset used in this study is the CREMA-D dataset. This dataset contains videos of actors expressing six types of emotions: happy, sad, angry, afraid, annoyed, and neutral. This dataset supports the development of voice- and image-based emotion recognition systems. To speed up the model training process, the dataset was processed using the Google platform

- Optimizer: Adam, Loss: Categorical Cross-entropy or Sparse Categorical.
- Regularization techniques: Dropout, EarlyStopping, ReduceLROnPlateau.

- Batch size and training time are adjusted based on GPU capacity; use Colab with GPU access after downloading from Kaggle.

B. Preprocessing

The datasets obtained in Section A were preprocessed. The preprocessing steps for each dataset are as follows: CREMA-D: OpenCV and a Haar cascade were used to extract faces from videos; the face images were resized to 64 x 64 pixels (for CNN) or 224 x 224 pixels (for the customized model). Audio is extracted using FFmpeg and converted into 40 x 128-dimensional MFCCs (Mel-Frequency Cepstral Coefficients). Normalization is performed to maintain input consistency.

C. Multimodal Feature Fusion

Multimodal feature fusion was performed by combining feature representations from visual and audio modalities to develop a multimodal emotion recognition system. In the visual branch, facial images were processed using a 2D CNN architecture to extract facial expression features, while in the audio branch, speech signals from the CREMA-D dataset were converted into 40×128 MFCC spectrograms and processed using a 2D CNN architecture to extract acoustic features. The extracted features from both modalities were subsequently combined at the feature fusion layer using the concatenate () method to form an integrated multimodal representation. The fused representation was then forwarded to the fully connected layer for the final emotion classification process.

D. Model Training

In this model training, the data used consists of facial and voice data obtained from the CREMA-D dataset. A multimodal Convolutional Neural Network (CNN) model was trained to classify human emotions based on the combination of facial expressions and voice features.

Before the training process was carried out, the initial step involved determining the storage location of the dataset to be used. The CREMA-D dataset was successfully uploaded and extracted into the Google Colab workspace, allowing all data to be accessed and processed by the system during the model training stage. The folder structure used consists of the main CREMA-D directory containing video data, audio data, and other supporting information required for the emotion classification process. The folder organization of the CREMA-D dataset includes the following directories:

- AudioMP3: These are .mp3 audio files that record the actor's vocal expressions based on specific emotions.
- AudioWAV: These contain the same audio files as AudioMP3 but in .wav format. This format is often used for audio signal processing because it is uncompressed and preserves the original quality.
- VideoFlash folder: This contains videos in .flv format that record both the actors' facial expressions and voices simultaneously. This allows for use in research involving the integration of visual and audio emotions.
- Docs folder: The docs folder may contain instructions or additional information about the dataset. This may include explanations of the file naming structure, emotion annotations, usage guidelines, and so on.
- ProcessedResults folder: The processedResults folder typically contains the results of processing or extracting raw data, such as feature extraction results, processed annotations, or format conversions.
- License.txt: The text file named License.txt contains the official license for the CREMA-D dataset. It is important to understand the restrictions on data use in research or publications.

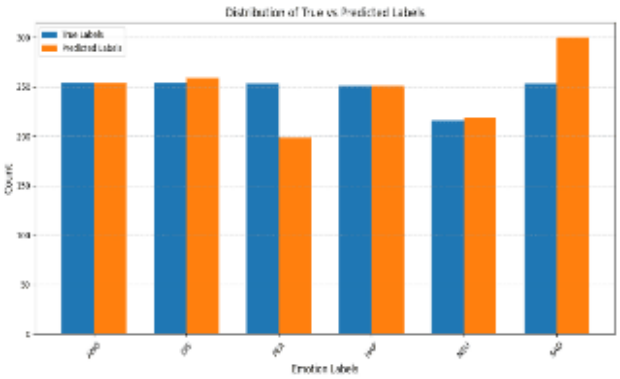


Fig. 5. Actual label distribution vs. model predictions on emotion data.

Fig. 5, the "True vs Predicted Distribution" chart, presents a comparison between the actual labels (True Labels) and the model predictions (Predicted Labels) across six emotion classes: ANG, DIS, FEA, HAP, NEU, and SAD. Overall, the true label distribution is relatively balanced, with most classes containing approximately 250 samples, while the NEU class has a slightly lower number of samples. The predicted distribution closely matches the true distribution for the ANG, DIS, HAP, and NEU classes, indicating that the model performs relatively well in recognizing these emotions. However, a noticeable discrepancy can be observed in the FEA class, where the number of predicted samples is significantly lower than the actual distribution, suggesting underprediction for this emotion category. In contrast, the SAD class shows a considerably higher number of predicted samples compared to the true labels, indicating that the model tends to over-predict the SAD emotion. Overall, the chart demonstrates that the model is capable of capturing the general distribution of most emotion classes, although some prediction bias remains, particularly between the FEA and SAD classes.

TABLE II. ACTUAL AND PREDICTED NUMBER OF LABELS BY EMOTION CLASS

Emotions Class	Number of Actual Data Points (True Labels)	Number of Model Predictions (Predicted Labels)	Difference (Predicted vs. Actual)
ANG (Angry)	255	255	0
DIS (Disgust)	255	260	+5
FEA (Fear)	255	200	-55
HAP (Happy)	250	250	0
NEU (Neutral)	215	220	+5
SAD (Sad)	255	300	+45

Table II presents the comparison between the number of actual data points (true labels) and the number of model predictions (predicted labels) for each emotion class. Overall, the model demonstrates relatively balanced prediction performance for several classes, particularly ANG and HAP, where the number of predicted samples exactly matches the actual data distribution. The DIS and NEU classes show only minor differences, with slight over-predictions of +5 samples each, indicating stable classification performance. However, significant discrepancies can be observed in the FEA and SAD classes. The FEA class is notably underpredicted by 55 samples, suggesting that the model has difficulty correctly identifying fear-related emotions. In contrast, the SAD class is over-predicted by 45 samples, indicating a tendency of the model to classify several other emotions as sadness. These results suggest that while the model is capable of maintaining good prediction consistency for some emotion categories, challenges still remain in distinguishing emotions with similar facial and acoustic characteristics, particularly between FEA and SAD.

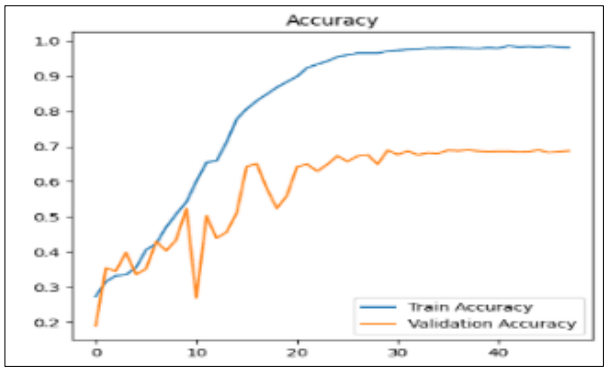


Fig. 6. Accuracy over epochs.

Fig. 6 illustrates the training and validation accuracy curves of the proposed multimodal CNN model during the training process. The training accuracy shows a consistent upward trend, eventually reaching approximately 98%, indicating that the model is capable of learning the training data effectively. Meanwhile, the validation accuracy also improves over time and stabilizes around 68–69%, although several fluctuations can be observed during the early and middle training epochs. The gap between training and validation accuracy suggests the presence of moderate overfitting; however, the validation performance remains relatively stable in the later epochs, indicating that the model still maintains reasonable generalization capability on unseen data.

Fig. 7 illustrates the training loss and validation loss curves of the proposed multimodal CNN model during the training process. The training loss decreases consistently and approaches zero, indicating that the model is able to learn the training data effectively. Meanwhile, the validation loss initially decreases but later exhibits noticeable fluctuations and tends to stabilize around 1.3–1.5 during the final epochs. The considerable gap between training loss and validation loss suggests the presence of overfitting, where the model performs very well on the training data but shows limited performance on the validation data. Nevertheless, the relatively stable validation loss in the later epochs indicates that the model still maintains a reasonable generalization capability on unseen data.

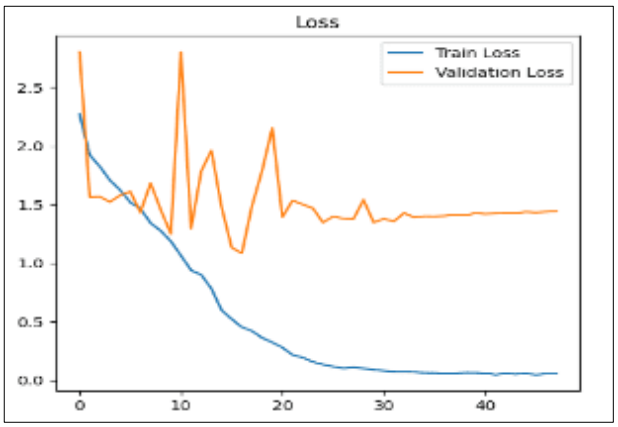
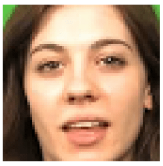
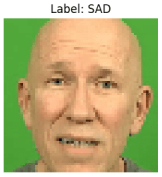


Fig. 7. Loss over epochs.

The video file format used in this study is .flv (Flash Video). This can be observed from the following code snippet: video_files = [f for f in os.listdir(video_dir) if f.endswith('.flv')], which indicates that the system only processes video files with the .flv extension. The .flv format was selected because it has a relatively small file size and is compatible with various video processing applications, thereby supporting efficiency in the video extraction and analysis processes in this study. Examples of facial and audio prediction results generated by the model used in this study are presented in Table III.

TABLE III. EXAMPLE OF SIX FACES AND AUDIO IMAGE PREDICTION

Label: NEU  Sample ke-1 — 1003_ITH_NEU_XX.flv	Label: FEA  Sample ke-2 — 1036_TSI_FEA_XX.flv	Label: HAP  Sample ke-3 — 1012_WSI_HAP_XX.flv
Label: DIS  Sample ke-4 — 1041_IOM_DIS_XX.flv	Label: SAD  Sample ke-5 — 1062_DFA_SAD_XX.flv	Label: DIS  Sample ke-6 — 1076_IWL_DIS_XX.flv

As shown in Table III, the model is capable of predicting emotions based on the combination of facial and audio data from several video samples.

The classification report generated from the evaluation of the multimodal (face and audio) emotion classification model is shown below. This report includes key evaluation metrics for each emotion class tested—namely, accuracy, recall, and F1-score as well as a summary of the model's overall performance.

Table IV presents the classification performance across six emotion classes using precision, recall, F1-score, and support (number of samples per class). The model performs best on the ANG class, achieving the highest F1-score of 0.82 with strong precision (0.82) and recall (0.82), indicating reliable and

balanced predictions. The HAP class also shows relatively good performance with an F1-score of 0.81. In contrast, DIS and FEA have the lowest F1-scores (0.67 and 0.58), reflecting the model's difficulty in accurately distinguishing these emotions. The SAD class exhibits high recall (0.61) but lower precision (0.66), suggesting that while many SAD instances are correctly identified, the model tends to overpredict this class. The NEU class shows moderate and balanced performance with an F1-score of 0.66. Overall, the model achieves an accuracy of 0.69 across 1,481 samples, with both macro and weighted averages also at 0.69, indicating consistent but still moderate performance and highlighting the need for further improvements in distinguishing similar emotional expressions.

TABLE IV. CLASSIFICATION REPORT

	Precision	Recall	F1-Score	support
ANG	0.82	0.82	0.82	254
DIS	0.67	0.68	0.67	254
FEA	0.66	0.51	0.58	253
HAP	0.81	0.81	0.81	251
NEU	0.66	0.67	0.66	216
SAD	0.56	0.66	0.61	253
Accuracy			0.69	1481
Macro avg	0.69	0.69	0.69	1481
Weighted avg	0.70	0.69	0.69	1481

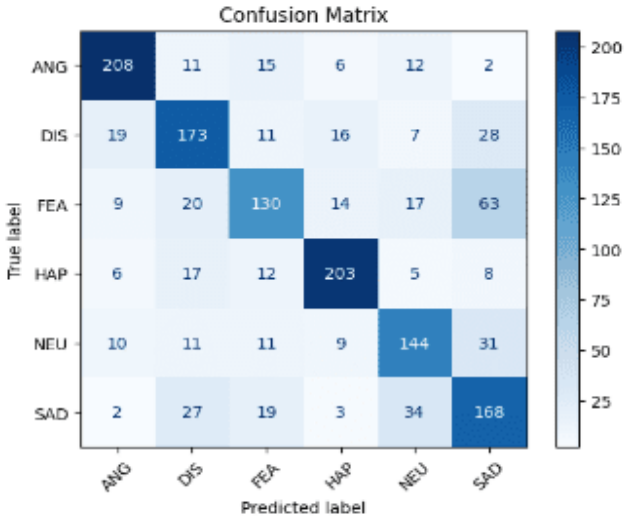


Fig. 8. Confusion matrix of emotion classification results using a multimodal CNN model.

Fig. 8 shows the confusion matrix illustrating the performance of the emotion classification model across six emotion classes, namely ANG, DIS, FEA, HAP, NEU, and SAD, where correct predictions are represented along the diagonal of the matrix. The model demonstrates the best performance on the ANG class with 208 correct predictions, followed by HAP with 203 correct predictions, DIS with 173 correct predictions, and SAD with 168 correct predictions, indicating strong recognition capability for these emotion categories. The NEU class also shows relatively good performance with 144 correctly classified samples, while the FEA class has the lowest performance with 130 correct predictions. Overall, the confusion matrix indicates that the proposed model achieves good classification performance across most emotion categories; however, it still faces challenges in distinguishing several negative emotions with similar characteristics, leading to misclassifications between classes. Visualization of Multimodal Emotion Classification Results with Confidence Scores.

TABLE V. VISUALIZATION OF MULTIMODAL EMOTION CLASSIFICATION RESULTS WITH CONFIDENCE SCORES

Class Label	Class Acc (%)	Audio Means (%)	Confidence (%)
ANG 	81.89	-0.69	92.48
DIS 	68.11	-0.17	99.91
FEA 	51.38	-0.26	83.57
HAP 	80.88	-0.58	99.99
NEU 	0.66	0.67	0.66
SAD 	0.56	0.66	0.61

Table V presents the visualization results of facial emotion classification using the proposed multimodal CNN model across six emotion classes, namely angry (ANG), disgust (DIS), fear (FEA), happy (HAP), neutral (NEU), and sad (SAD). Each facial sample is accompanied by class accuracy, audio mean,

and confidence score values, which indicate the model's confidence level in performing the prediction. The model achieves the best performance on the ANG class with an accuracy of 81.89% and the HAP class with an accuracy of 80.88%, while the FEA class shows the lowest performance with an accuracy of 51.38%, indicating that fear emotion remains difficult to distinguish from other emotions. The confidence scores for most classes are above 90%, demonstrating that the model has a high level of confidence in its classification results. In addition, the audio mean values represent the average contribution of audio features for each emotion prediction. Overall, this visualization demonstrates that the multimodal model is capable of recognizing most emotions effectively through the combination of visual and audio information, although challenges still remain in classifying emotions with similar facial expression characteristics.

IV. Conclusion

The results of this study demonstrate that the proposed multimodal CNN model is capable of effectively recognizing human emotions by integrating facial image and audio features. Based on the evaluation results, the model achieved an overall accuracy of 69%, with weighted precision, recall, and F1-score values of 0.70, 0.69, and 0.69, respectively. Among the evaluated emotion classes, the ANG (Angry) class achieved the best performance with an F1-score of 0.82, followed by the HAP (Happy) class with an F1-score of 0.81, indicating that the model was highly effective in recognizing these emotions. In contrast, the FEA (Fear) class showed the lowest performance with an F1-score of 0.58, suggesting that fear-related emotions remain challenging to distinguish due to similarities with other negative emotional expressions. The experimental results further indicate that the proposed multimodal approach was able to integrate visual and audio information effectively, resulting in stable performance across several major emotion classes. In addition, the developed CNN architecture provides an efficient and practical framework for multimodal emotion recognition through feature fusion between facial expressions and speech signals. Overall, this study confirms that combining facial and audio modalities can improve emotion classification performance and enhance the model's generalization capability on diverse testing data. Future work may focus on improving recognition performance for challenging emotion classes and incorporating additional modalities or advanced fusion strategies to further enhance classification accuracy.

Acknowledgment

The authors would like to express their sincere gratitude to all parties who have provided support, assistance, and contributions throughout the completion of this research. Special appreciation is extended to the authors and their families for their prayers, motivation, and moral support. The authors also gratefully acknowledge Universitas Sriwijaya, Palembang, for providing academic facilities, a supportive research environment, and institutional support that made this study possible.

References

[1] C. E. Valderrama, M. G. Gomes Ferreira, J. M. Mayor Torres, A. R. Garcia-Ramirez, and S. G. Camorlinga, "Editorial: Machine learning approaches to recognize human emotions," Front. Psychol., vol. 14, 2023, doi: 10.3389/fpsyg. 2023.1333794.

[2] C. Qiana, J. A. L. Marquesb, and A. R. de Alexandria, "Real-time emotion recognition based on facial expressions using Artificial Intelligence techniques: A review and future directions," Multidiscip. Rev., vol. 8, no. 10, pp. 1–27, 2025, doi: 10.31893/multirev. 2025328.

[3] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," 2025. doi: 10.3390/biomimetics10070418.

[4] R. Pereira et al., "Systematic Review of Emotion Detection with Computer Vision and Deep Learning," 2024. doi: 10.3390/s24113484.

[5] T. Chutia and N. Baruah, "A review on emotion detection by using deep learning techniques," Artif. Intell. Rev., vol. 57, no. 8, p. 203, 2024, doi: 10.1007/s10462-024-10831-1.

[6] M. Karthiga, E. Suganya, S. Sountharrajan, B. Balusamy, and S. Selvarajan, "EEG-based smart emotion recognition using meta heuristic optimization and hybrid deep learning techniques," Sci. Rep., vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-80448-5.

[7] X. Wang, Y. Ren, Z. Luo, W. He, J. Hong, and Y. Huang, "Deep learning-based EEG emotion recognition: Current trends and future perspectives.," Front. Psychol., vol. 14, p. 1126994, 2023, doi: 10.3389/fpsyg. 2023.1126994.

[8] E. Gkintoni, A. Aroutzidis, H. Antonopoulou, and C. Halkiopoulos, "From Neural Networks to Emotional Networks: A Systematic Review of EEG-Based Emotion Recognition in Cognitive Neuroscience and Real-World Applications.," Brain Sci., vol. 15, no. 3, Feb. 2025, doi: 10.3390/brainsci15030220.

[9] A. M. S. Gonzalez-Acosta et al., "The first look: a biometric analysis of emotion recognition using key facial features," Front. Comput. Sci., vol. 7, 2025, doi: 10.3389/fcomp. 2025.1554320.

[10] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using a convolutional neural network on the Emotion Detection Dataset," Sci. Rep., vol. 14, no. 1, p. 14429, 2024, doi: 10.1038/s41598-024-65276-x.

[11] S. A. Salloum, K. M. Alomari, A. M. Alfaisal, R. A. Aljanada, and A. Basiouni, "Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods," Smart Learn. Environ., vol. 12, no. 1, p. 21, 2025, doi: 10.1186/s40561-025-00374-5.

[12] T.-W. Kim and K.-C. Kwak, "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques," 2024. doi: 10.3390/app14041553.

[13] H. Alharbi, "Explainable feature selection and deep learning-based emotion recognition in virtual reality using eye tracker and physiological data," Front. Med., vol. 11, p. 1438720, 2024, doi: 10.3389/fmed.2024.1438720.

[14] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human Emotion Recognition: Review of Sensors and Methods.," Sensors (Basel)., vol. 20, no. 3, Jan. 2020, doi: 10.3390/s20030592.

[15] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology - a systematic literature review.," BMC Psychol., vol. 12, no. 1, p. 95, Feb. 2024, doi: 10.1186/s40359-024-01581-4.

[16] S. M. George and P. Muhamed Ilyas, "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise," Neurocomputing, vol. 568, p. 127015, 2024, doi: https://doi.org/10.1016/j.neucom.2023.127015.

[17] A. Pentari, G. Kafentzis, and M. Tsiknakis, "Speech emotion recognition via graph-based representations," Sci. Rep., vol. 14, no. 1, p. 4484, 2024, doi: 10.1038/s41598-024-52989-2.

[18] G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," EURASIP J. Audio, Speech, Music Process., vol. 2023, no. 1, p. 22, 2023, doi: 10.1186/s13636-023-00289-4.

[19] A. Chakhtouna, S. Sekkate, and A. Adib, "Efficient bimodal emotion recognition system based on speech/text embeddings and ensemble learning fusion," Ann. Telecommun., vol. 80, no. 5, pp. 379–399, 2025, doi: 10.1007/s12243-025-01088-y.

[20] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," Artif. Intell. Rev., vol. 58, no. 2, p. 49, 2024, doi: 10.1007/s10462-024-11065-x.

[21] M. Kaur and M. Kumar, "Facial emotion recognition: A comprehensive review," Expert Syst., vol. 41, no. 10, p. e13670, 2024.

[22] O. Kalyta, O. Barmak, P. Radiuk, and I. Krak, "Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics," 2023. doi: 10.3390/app13179890.

[23] Z.-Y. Huang et al., "A study on computer vision for facial emotion recognition," Sci. Rep., vol. 13, no. 1, p. 8425, 2023, doi: 10.1038/s41598-023-35446-4.

[24] B. Fang, Y. Zhao, G. Han, and J. He, "Expression-Guided Deep Joint Learning for Facial Expression Recognition," Sensors (Basel)., vol. 23, no. 16, Aug. 2023, doi: 10.3390/s23167148.

[25] A. Talukder and S. Ghosh, "Facial Image expression recognition and prediction system," Sci. Rep., vol. 14, no. 1, p. 27760, 2024, doi: 10.1038/s41598-024-79146-z.

[26] P. Thakur, N. Kaur, N. Aggarwal, and S. Singh, "A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations," Expert Syst., vol. 42, no. 9, p. e70103, 2025.

[27] S. Gutiérrez, J. Fernández-Novales, T. Garde-Cerdán, S. Marín-San Román, J. Tardaguila, and M. P. Diago, "Multi-sensor spectral fusion to model grape composition using deep learning," Inf. Fusion, vol. 99, p. 101865, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101865.

[28] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: a comprehensive review," Neural Comput. Appl., vol. 36, no. 16, pp. 8901–8947, 2024, doi: 10.1007/s00521-024-09426-2.

[29] J.-H. Lee, J.-Y. Kim, and H.-G. Kim, "Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning," Bioeng. (Basel, Switzerland), vol. 11, no. 10, Oct. 2024, doi: 10.3390/bioengineering11100997.

[30] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," Int. J. Inf. Technol., vol. 15, no. 4, pp. 1777–1787, 2023, doi: 10.1007/s41870-023-01184-z.

[31] R. Baazeem, "Explainable cross-domain emotion recognition using non-linear optimization and multimodal feature fusion based deep learning model," Int. J. Inf. Technol., 2025, doi: 10.1007/s41870-025-02930-1.

[32] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," Sci. Rep., vol. 14, no. 1, p. 13126, 2024.

[33] P. Koromilas and T. Giannakopoulos, "Deep Multimodal Emotion Recognition on Human Speech: A Review," 2021. doi: 10.3390/app11177962.

[34] B. Akande, "Multimodal Deep Learning for Emotion Detection from Text, Audio, and Facial Expressions," no. April, 2025.

[35] A.-L. Cîrneanu, D. Popescu, and D. Iordache, "New Trends in Emotion Recognition Using Image Analysis by Neural Networks, A Systematic Review," 2023. doi: 10.3390/s23167092.

[36] A. R. Khan, "Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges," 2022. doi: 10.3390/info13060268.

[37] S. Son and Y. Jeong, "Face and Voice Recognition-Based Emotion Analysis System (EAS) to Minimize Heterogeneity in the Metaverse," Appl. Sci., vol. 15, no. 2, 2025, doi: 10.3390/app15020845.

[38] B. T. Atmaja and A. Sasou, "Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations," 2022. doi: 10.3390/s22176369.

[39] Y. Zhao and X. Shu, "Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC)," Sci. Rep., vol. 13, no. 1, p. 20398, 2023.

[40] C. Xu, Y. Liu, W. Song, Z. Liang, and X. Chen, "A New Network Structure for Speech Emotion Recognition Research," Sensors (Basel)., vol. 24, no. 5, Feb. 2024, doi: 10.3390/s24051429.

[41] D. Resende Faria, A. I. Weinberg, and P. P. Ayrosa, "Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning," 2024. doi: 10.3390/app14156631.

[42] K. Kaur and P. Singh, "Trends in speech emotion recognition: a comprehensive survey," Multimed. Tools Appl., vol. 82, no. 19, pp. 29307–29351, 2023.

[43] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," Entropy (Basel), vol. 25, no. 10, Oct. 2023, doi: 10.3390/e25101440.

[44] Y. Zeng, J.-W. Zhang, and J. Yang, "Multimodal emotion recognition in the metaverse era: New needs and transformation in mental health work," World J. Clin. cases, vol. 12, no. 34, p. 6674, 2024.

[45] A. A. Wafa, M. M. Eldefrawi, and M. S. Farhan, "Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning," J. Big Data, vol. 12, no. 1, p. 210, 2025.

[46] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion," Sci. Rep., vol. 15, no. 1, p. 5473, 2025.

[47] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019-Septe, pp. 2828–2832, 2019, doi: 10.21437/Interspeech.2019-2293.

[48] N. Saleem et al., "DeepCNN: Spectro-temporal feature representation for speech emotion recognition," CAAI Trans. Intell. Technol., vol. 8, no. 2, pp. 401–417, 2023, doi: 10.1049/cit2.12233.

[49] L. Wijayasingha and J. A. Stankovic, "Robustness to noise for speech emotion classification using CNNs and attention mechanisms," Smart Heal., vol. 19, p. 100165, 2021, doi: https://doi.org/10.1016/j.smhl.2020.100165.

[50] M. F. Almufareh, S. Kausar, M. Humayun, and S. Tehsin, "A Conceptual Model for Inclusive Technology: Advancing Disability Inclusion through Artificial Intelligence," J. Disabil. Res., vol. 3, no. 1, pp. 1–11, 2024, doi: 10.57197/JDR-2023-0060.

[51] R. Malviya and S. Rajput, "AI-Driven Innovations in Assistive Technology for People with Disabilities BT - Advances and Insights into AI-Created Disability Supports," R. Malviya and S. Rajput, Eds., Singapore: Springer Nature Singapore, 2025, pp. 61–77. doi: 10.1007/978-981-96-6069-8_4.

[52] N. Simić et al., "Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks," 2024. doi: 10.3390/app14041325.

[53] F. Makhmudov, A. Kultimuratov, and Y.-I. Cho, "Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures," 2024. doi: 10.3390/app14104199.

[54] S. Garcia, F. Gomez-Donoso, and M. Cazorla, "Enhancing Human–Robot Interaction: Development of Multimodal Robotic Assistant for User Emotion Recognition," 2024. doi: 10.3390/app142411914.

[55] K. Chemnad, "Digital accessibility in the era of artificial intelligence — Bibliometric analysis and systematic review," Front. Artif. Intell., 2022.

[56] M. M. Terras, D. Jarrett, and S. A. McGregor, "The Importance of Accessible Information in Promoting the Inclusion of People with an Intellectual Disability," 2021. doi: 10.3390/disabilities1030011.

[57] H. M. Salman and R. T. Rasheed, "Smart Door for Handicapped People via Face Recognition and Voice Command Technique," Eng. Technol. J., vol. 39, no. 1B, pp. 222–230, 2021, doi: 10.30684/etj.v39i1b.1719.

[58] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," Nat. Med., vol. 28, no. 9, pp. 1773–1784, 2022, doi: 10.1038/s41591-022-01981-2.

[59] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," Int. J. Appl. Earth Obs. Geoinf., vol. 112, p. 102926, 2022, doi: https://doi.org/10.1016/j.jag.2022.102926.

[60] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on Bengali social media comments using machine learning," Int. J. Cogn. Comput. Eng., vol. 4, pp. 21–35, 2023, doi: https://doi.org/10.1016/j.ijcce.2023.01.001.

[61] W. Chai and G. Wang, "Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend," 2022. doi: 10.3390/app12136588.

[62] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Inf. Fusion, vol. 91, pp. 424–444, Mar. 2023, doi: 10.1016/j.inffus.2022.09.025.

[63] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision : advances, trends, applications, and datasets," Vis. Comput., vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[64] A. Mukherjee and M. Shrivastava, "Lost in Translation? Found in Evaluation: A Comprehensive Survey on Sentence-Level Translation Evaluation," ACM Comput. Surv., vol. 58, no. 1, Sep. 2025, doi: 10.1145/3735970.

[65] P. J. Vaz, J. M. F. Rodrigues, and P. J. S. Cardoso, "Affective Computing Emotional Body Gesture Recognition: Evolution and the Cream of the Crop," IEEE Access, vol. 13, no. September, pp. 192871–192890, 2025, doi: 10.1109/ACCESS.2025.3630563.

[66] S. Learn, "scikit-learn Machine Learning in Python," scikit-learn. Accessed: Jul. 02, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?utm_source=chatgpt.com#

[67] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," Technometrics, vol. 64, no. 2, pp. 166–176, Apr. 2022, doi: 10.1080/00401706.2021.1921037.

[68] J. Park, E. Byon, Y. M. Ko, and S. Shashaani, "Strata design for variance reduction in stochastic simulation," Technometrics, vol. 67, no. 2, pp. 203–214, 2025.

[69] H. Liu and M. Cocea, "Semi-random partitioning of data into training and test sets in granular computing context," Granul. Comput., vol. 2, no. 4, pp. 357–386, 2017, doi: 10.1007/s41066-017-0049-2.

[70] O. Yazdanpanah, A. Formisano, M. Chang, and B. Mohebi, "Fragility curves for seismic damage assessment in regular and irregular MRFs using improved wavelet-based damage index," Measurement, vol. 182, p. 109558, 2021, doi: https://doi.org/10.1016/j.measurement.2021.109558.

[71] R. Cantini et al., "Block size estimation for data partitioning in HPC applications using machine learning techniques," J. Big Data, vol. 11, no. 1, p. 19, 2024, doi: 10.1186/s40537-023-00862-w.

[72] O. Elharrouss et al., "Task-based Loss Functions in Computer Vision: A Comprehensive Review," Apr. 01, 2025. doi: 10.48550/arXiv.2504.04242.

[73] J.-S. Hwang, S.-S. Lee, J.-W. Gil, and C.-K. Lee, "Determination of Optimal Batch Size of Deep Learning Models with Time Series Data," 2024. doi: 10.3390/su16145936.

[74] J. Liu, H. Wang, M. Sun, and Y. Wei, "Graph based emotion recognition with attention pooling for variable-length utterances," Neurocomputing, vol. 496, pp. 46–55, 2022, doi: https://doi.org/10.1016/j.neucom.2022.05.007.

[75] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light-weight deep neural ensemble model using handcrafted features," Sci. Rep., vol. 15, no. 1, p. 11824, 2025, doi: 10.1038/s41598-025-95734-z.

[76] Y. Ahn, S. Han, S. Lee, and J. W. Shin, "Speech Emotion Recognition Incorporating Relative Difficulty and Labeling Reliability," Sensors (Basel)., vol. 24, no. 13, Jun. 2024, doi: 10.3390/s24134111.

[77] W. J. C, G. J. W. Kathrine, V. Shanmuganathan, and S. Sumathi, "Improved optimizer with deep learning model for emotion detection and classification," vol. 21, no. June, pp. 6631–6657, 2024, doi: 10.3934/mbe.2024290.

[78] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," Multimed. Tools Appl., vol. 84, no. 21, pp. 24079–24190, 2025, doi: 10.1007/s11042-024-20011-6.

[79] M. Abdullahi et al., "A systematic literature review of visual feature learning: deep learning techniques, applications, challenges and future directions," Multimed. Tools Appl., vol. 84, no. 19, pp. 20439–20496, 2025, doi: 10.1007/s11042-024-19823-3.

[80] P. Amalia et al., "Implementasi Algoritma Convolutional Neural Network (CNN) Dengan Optimizer Adam Dalam Deteksi Emosional Pada Wajah Manusia," J. VOKASI Tek., vol. 3, no. 1, pp. 13–22, 2025, [Online]. Available: https://mentech.id/jurnal/index.php/juvotek/article/view/60

[81] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning.," Sci. Rep., vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[82] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification.," PLoS One, vol. 19, no. 12, p. e0316019, 2024, doi: 10.1371/journal.pone.0316019.

[83] P. Mishra, A. S. Verma, P. Chaudhary, and A. Dutta, "Emotion Recognition from Facial Expression Using Deep Learning Techniques," 2024 IEEE 9th Int. Conf. Converg. Technol. I2CT 2024, 2024, doi: 10.1109/I2CT61223.2024.10543313.

[84] D. Resha, P. Pamungkas, and B. H. Prasetio, "Implementasi Ekstraksi Gammatone-Frequency Cepstral Coefficient dan Klasifikasi Hidden Markov Model dalam Identifikasi Emosi Menggunakan Suara Jantung," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 9, no. 3, pp. 1–10, 2025, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14533

[85] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," 2023. doi: 10.3390/computation11030052.

[86] A. Al Bataineh, D. Kaur, M. Al-khassaweneh, and E. Al-sharoa, "Automated CNN Architectural Design: A Simple and Efficient Methodology for Computer Vision Tasks," 2023. doi: 10.3390/math11051141.

[87] G. Ortega-Flores, G. Altamirano-Escobedo, D. Mercado-Ravell, and E. Bayro-Corrochano, "Quaternion CNN in Deep Learning Processing for EEG with Applications to Brain Disease Detection," 2025. doi: 10.3390/app152111526.

[88] F. A. Albalooshi and M. R. Qader, "Deep Learning Algorithm for Automatic Classification of Power Quality Disturbances," 2025. doi: 10.3390/app15031442.

[89] G. Hernández-Nava, S. Salazar-Colores, E. Cabal-Yepez, and J.-M. Ramos-Arreguín, "Parallel Ictal-Net, a Parallel CNN Architecture with Efficient Channel Attention for Seizure Detection," 2024. doi: 10.3390/s24030716.

[90] I. Rakhmatulin, M.-S. Dao, A. Nassibi, and D. Mandic, "Exploring Convolutional Neural Network Architectures for EEG Feature Extraction," 2024. doi: 10.3390/s24030877.

[91] R. Begazo, A. Aguilera, I. Dongo, and Y. Cardinale, "A Combined CNN Architecture for Speech Emotion Recognition," 2024. doi: 10.3390/s24175797.

[92] X. Liu, C. Cao, and S. Duan, "A Low-Power Hardware Architecture for Real-Time CNN Computing," 2023. doi: 10.3390/s23042045.

[93] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[94] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," 2024. doi: 10.3390/info15120755.

Interlayer Coupling Strategy for Two-Layer Interdependent Command-and-Control Networks: Balancing Node Walk Betweenness and Operational Attributes

Bo Chen[1], Yangkun Wang[2], Yufeng Chen[3]*, Yuqi Yang[4]
College of Electronic and Electrical Engineering, Lingnan Normal University, Zhanjiang, China[1]
School of Robotics and Automation, Hubei University of Automotive Technology, Shiyan, China[2]
School of Intelligent Connected Vehicle, Hubei University of Automotive Technology, Shiyan, China[3, 4]

Abstract*—In system-of-systems operations, command-and-control networks (C2Ns) increasingly exhibit hierarchical interdependence and cross-network coupling. Designing effective interlayer coupling strategies is therefore essential for improving the survivability of interdependent C2Ns under attacks. However, most existing coupling strategies are predominantly topology-driven and fail to jointly capture multipath information transmission characteristics and operational compatibility among heterogeneous nodes. To address this issue, this study proposes an interlayer coupling strategy for two-layer interdependent C2Ns, referred to as PRO, by integrating node walk betweenness and operational attributes. Specifically, a walk-betweenness-based dependency strength metric is constructed to characterize node importance from the perspective of multipath information transmission, while an operational attribute matching mechanism is introduced to evaluate the functional compatibility of cross-layer nodes. Based on these two components, the PRO strategy is developed for survivability-oriented interlayer coupling design. Simulation experiments on an air-defense command-and-control scenario are conducted under both deliberate and random attacks. The results of repeated simulations show that the proposed strategy consistently outperforms the compared baseline methods in node survival rate, observe-orient-decide-act (OODA)-oriented functional chain retention ratio, and combat link efficiency degradation rate. Compared with the baseline random strategy, under random attack, PRO improves node survival rate by 50.00%, improves the OODA-oriented functional chain retention ratio by 19.40%, and reduces the combat link efficiency degradation rate by 3.53%. These results indicate that the PRO strategy improves not only structural robustness but also the preservation of sensing–command–firepower coordination under cascading failures. The proposed method provides a joint structural–functional framework for mitigating cascading failure risks and enhancing survivability in interdependent C2Ns.

Keywords*—Interdependent command and control networks; interlayer coupling strategy; walk betweenness; operational attributes

I. Introduction

In system-of-systems operations, command-and-control networks (C2Ns) increasingly exhibit hierarchical interdependence and cross-network coupling. Sensing, command-and-control, and firepower subnetworks interact through information sharing, command dissemination, strike execution, and feedback, thereby forming an observe-orient-decide-act (OODA) process [1], [19]. Because different combat units are functionally heterogeneous and strongly interdependent, local failures may propagate across subnetworks through dependency relationships, eventually degrading not only structural connectivity but also mission-oriented coordination capability. Therefore, introducing the perspective of interdependent networks is important for understanding the survivability of C2Ns, while the design of interlayer dependency relationships becomes a key factor affecting cascading-failure propagation and mission-oriented coordination.

Existing studies have shown that coupling patterns and coupling preferences significantly affect the robustness of interdependent networks. However, command-and-control networks differ from generic interdependent networks because they contain heterogeneous operational units, hierarchical command relationships, and mission-oriented sensing–command–firepower coordination. As reviewed in Section II, existing coupling strategies are still insufficient for interdependent C2Ns because they usually rely on topology-driven criteria and rarely integrate multipath information transmission characteristics with operational compatibility among heterogeneous nodes.

To address this gap, this study proposes an interlayer coupling strategy for two-layer interdependent C2Ns, referred to as PRO. The proposed strategy integrates walk-betweenness-based dependency strength and operational attribute matching. The former characterizes node importance from the perspective of multipath information transmission, whereas the latter evaluates the functional compatibility of cross-layer heterogeneous nodes. In this way, PRO supports survivability-oriented interlayer dependency design from a joint structural–functional perspective. The main contributions of this study are as follows:

- A dependency strength metric was proposed based on node walk betweenness, considering the multipath characteristics of node information transmission.

*Corresponding author.

This work was supported by the Project of Natural Science Foundation of Guangdong Province (2024A1515011674,2023A1515011568) and Open Experiment Project of CENI (China Environment for Network Innovations), China Major Scientific Infrastructure (2022A0213).

- A node operational attribute matching metric was established to characterize the functional compatibility among heterogeneous combat units.
- An interlayer coupling strategy for two-layer interdependent C2Ns was developed by integrating dependency strength and operational attribute matching.
- Simulation experiments under deliberate and random attacks are conducted to evaluate the proposed strategy from the perspectives of node survival rate, OODA-oriented functional chain retention ratio, and combat link efficiency degradation rate, thereby verifying its structural and functional survivability advantages.

The remainder of this study is organized as follows. Section II reviews related studies on interdependent-network coupling strategies and C2N survivability analysis and clarifies the research gap. Section III abstracts the two-layer interdependent C2N model. Section IV defines the walk-betweenness-based dependency strength, operational attribute matching metric, and the proposed PRO coupling strategy. Section V presents simulation experiments and comparative results. Section VI discusses the mechanisms, practical applicability, and scalability of the proposed method. Section VII presents future research directions. Finally, Section VIII concludes the study.

II. Related Work

A. Coupling Strategies in Interdependent Networks

In recent years, studies have reported that the robustness of interdependent networks is closely related to the choice of coupling strategies [2]-[4]. Existing coupling strategies can be broadly divided into coupling-pattern-based methods and topology-preference-based methods. Coupling-pattern-based methods mainly focus on how dependency edges are organized between networks. For example, Xu and Fu [5] compared one-to-one, one-to-many, and many-to-many coupling patterns in directed-undirected interdependent networks and analyzed their effects on cascading-failure robustness. Liu et al. [6] compared assortative, disassortative, and random coupling strategies. Chen and Du et al. [7] further showed that different coupling preferences may lead to different robustness behaviors under sparse and dense coupling conditions.

Topology-preference-based methods further optimize interlayer dependency relationships according to the structural properties of nodes or edges. Hao et al. [8] proposed assortative and disassortative coupling strategies based on node proximity and shortest-path information. Wang et al. [9] proposed a locally optimized coupling strategy based on neighbor node priority connection (NPC), and Zang et al. [10] further developed a low-degree neighbor node priority connection (LDNPC) strategy. Li et al. [11] investigated edge-coupled interdependent networks under different coupling modes. Lin et al. [12] introduced a stepwise optimization strategy to dynamically adjust coupling coefficients during cascading failures. Gao et al. [13] proposed a globally homogeneous coupling strategy based on node degree and average network degree. Ji et al. [14] defined inter-degree differences and proposed random inter-degree–degree difference (RIDD) and low inter-degree–degree difference (LIDD) strategies.

These studies demonstrate that coupling patterns, coupling preferences, and topology-based optimization rules can significantly affect the robustness of interdependent networks. However, most of these strategies are developed for generic interdependent networks. Their coupling criteria are mainly derived from static topology, local connectivity, shortest-path proximity, or degree-related indicators. Therefore, they are insufficient for command-and-control networks, where interlayer dependency relationships not only represent structural connections but also reflect information transmission, command support, and operational coordination among heterogeneous combat units.

B. Command-and-Control Network Modeling and Survivability Analysis

Compared with generic interdependent networks, C2Ns have stronger mission semantics, functional heterogeneity, and hierarchical organization. Sensing, command-and-control, and firepower nodes perform different operational roles and jointly support the observe–orient–decide–act (OODA) process. Therefore, survivability analysis of C2Ns should consider not only structural connectivity but also the preservation of mission-oriented sensing–command–firepower functional chains.

Existing studies have investigated C2Ns from different perspectives. Wang et al. [15] proposed an invulnerability measure for C2Ns based on mission links. Chen et al. [16] developed a dual-layer interdependent C2N model for integrated reconnaissance-strike and OODA-loop capabilities. Chen et al. [17] investigated the dynamic evolution of C2Ns using a hypernetwork-based model. Pan et al. [18] proposed a multi-attribute-weighted modeling method for C2Ns. Ling et al. [19] discussed network-centric warfare metrics from connectivity to the OODA cycle. Zhao et al. [20] further highlighted the importance of robustness analysis of C2Ns under cascading failures.

These studies provide important foundations for C2N modeling, mission-oriented performance evaluation, and cascading-failure analysis. Nevertheless, most of them focus on network abstraction, mission-link evaluation, dynamic evolution, or failure propagation mechanisms. The problem of how to design interlayer dependency edges from a joint structural-functional perspective remains insufficiently investigated.

C. Research Gap and Motivation

The above studies indicate that coupling design is critical for the robustness of interdependent networks and that C2Ns require mission-oriented modeling and evaluation. However, three limitations remain in existing interlayer coupling studies.

First, many coupling strategies rely on degree-related indicators, which mainly describe local connectivity and cannot fully capture the indirect influence of nodes in global information transmission. Second, shortest-path-based indicators such as betweenness centrality may be insufficient for C2Ns, because information transmission in C2Ns may follow multiple redundant, non-shortest, or dynamically selected paths.

Third, topology-driven coupling strategies do not explicitly evaluate whether cross-layer nodes are functionally compatible in terms of sensing, command support, and firepower coordination.

These limitations motivate the proposed PRO strategy. By introducing walk-betweenness-based dependency strength, PRO characterizes node importance from the perspective of multipath information transmission. By introducing operational attribute matching, PRO further evaluates the functional compatibility of heterogeneous cross-layer nodes. Therefore, the proposed strategy provides a joint structural-functional coupling framework for survivability-oriented interlayer dependency design in two-layer interdependent C2Ns.

III. Abstraction of Two-Layer Interdependent C2N Modeling

A. Characterization of C2N

With the continuous evolution of combat modes and battlefield environments, the relationships among various combat units within C2Ns have become increasingly complex and diverse. Therefore, C2Ns are characterized by node heterogeneity, interlaced links, and multi-network interdependence. To simplify model analysis, we adopted the two-layer interdependent C2N model abstracted in Ref. [16]. The model was constructed as follows: 1) The firepower and sensing subnetworks were independently constructed using an attribute synergy strategy. 2) The aforementioned subnetworks were then connected using a detect–strike integration edge strategy to form the upper-layer firepower–sensing fusion network; 3) An improved tree-like network structure was used to construct the command and control subnetwork, forming the lower-layer network; 4) An interlayer coupling strategy was employed to connect the upper and lower layers, resulting in a complete two-layer interdependent C2N. The specific structure of the two-layer interdependent C2N is shown in Fig. 1. In this structure, the upper layer is the firepower–sensing fusion network, formed by integrating the sensing and firepower subnetworks; the lower layer is the command-and-control subnetwork. Intra-layer edges represent synergistic relationships within a subnetwork, while interlayer edges represent dependency relationships between subnetworks.

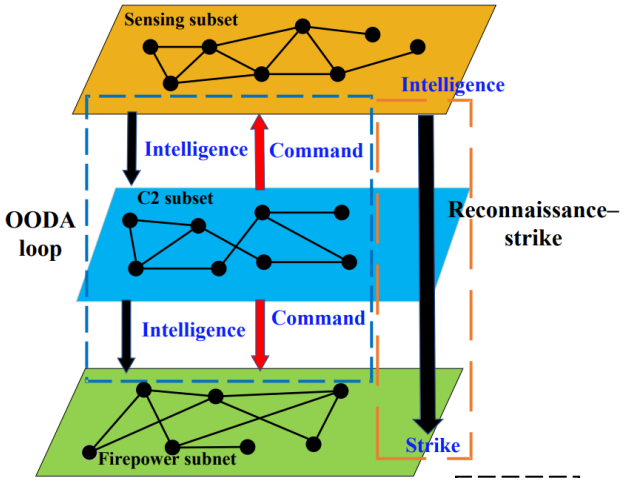


Fig. 1. Structure of the two-layer interdependent C2N.

This model can be represented by the adjacency matrix as follows:

$$W = \begin{bmatrix} W_{sf} & W_{sf-c} \\ W_{c-sf} & W_c \end{bmatrix} \quad (1)$$

where, W_{sf} is the upper-layer firepower–sensing fusion network; W_c is the lower-layer command and control subnetwork; W_{sf-c} and W_{c-sf} are the interdependent coupling networks between the two subnetworks.

B. Node Abstraction

Combat units in the C2N are abstracted as sensing nodes V_s, command and control nodes V_c, and firepower nodes V_f. Considering the operational attribute characteristics of each type of node, they are modeled as follows:

$$V(i) = \langle ID_V, T_V, H_V, Attribute_V, DependentNode_v \rangle \quad (2)$$

where, ID_V is the unique network sequence identifier of the node; T_V indicates the node type, including sensing, command and control, and firepower; H_V is the node's hierarchical level, with lower values indicating higher levels; $Attribute_V$ is the normalized attribute vector of the node, modeled as a scenario-based role-related descriptor. For sensing nodes, it mainly characterizes detection-related capability; for command-and-control nodes, it mainly characterizes communication and command-support capability; and for firepower nodes, it mainly characterizes strike-related capability. $DependentNode_v$ is the node(s) on which node i is dependent. To clarify the source and semantic meaning of the operational attributes used in this study, the attribute categories and representative components of sensing, command-and-control, and firepower nodes are summarized in Table I.

TABLE I. Operational Attribute Categories and Semantic Meanings of Heterogeneous Nodes in the C2N

Node type	Attribute semantic meaning	Representative attribute components
Sensing node	Detection-related capability	Detection range, detection accuracy, reconnaissance timeliness
Command-and-control node	Communication and command-support capability	Bandwidth, signal-to-noise ratio, hierarchical level
Firepower node	Strike-related capability	Strike range, firepower intensity, response timeliness

C. Link Abstraction

Combat units within the C2N exchange and process different types of information via wired, wireless, and other communication means. Edges can be represented as

$$E(m) = \left\langle ID_{E(m)}, type_{E(m)}, DependentAttribute_{E_{(m)}} \right\rangle \quad (3)$$

where, $ID_{E(m)}$ is the identifier of an interlayer edge; $type_{E(m)}$ is the edge type, including intra-layer synergistic and interlayer dependent edges; $DependentAttribute_{E_{(m)}}$ is the dependency strength of a dependent edge.

IV. Interlayer Coupling Strategy for Two-Layer Interdependent C2Ns

A. Node Walk Betweenness

C2Ns employ diverse communication methods, and information dissemination follows multipath transmission patterns. Taking into account the weights of different path lengths and based on the concept of betweenness, we introduced the notion of walk betweenness to characterize the frequency with which a node acts as an intermediary across all possible walk paths. Walk betweenness accounts for path length weighting. The core idea is to count the number of times a node appears in walks from every source to the destination, using a decay factor to handle the issue of infinite paths.

The walk betweenness of node v_i is defined as the proportion of the information volume ultimately received by that node in the total information volume of the network when each node in the network sequentially acts as a unique source to send an initial message (total information normalized to 1), and the message propagates according to a hierarchical walk rule. The calculation procedure is as follows:

Step 1: Initialize the information volume of each node in the C2N to I_0. Assume that in each information walk iteration, only one node sends information while others receive it. The initial information volume of a node is

$$H_0 = [I_0, I_0, \cdots, I_0]_N \tag{4}$$

where, N is the total number of nodes.

Step 2: During the walk process, the information volume of node v_j is evenly distributed to its neighboring nodes, and its own information volume is reset to zero. If the degree of node v_j is k_j, then after n iterations, the information volume received by a directly connected node v_i is

$$H_n(v_i) = H_{n-1}(v_j)/I_0 g k_j \tag{5}$$

where, $i \neq j$ and $H_{n-1}(v_j)$ are the information volume possessed by node v_j after the $(n\text{-}1)$th iteration. The total number of iterations does not exceed the network hierarchy level D, i.e., $n = 0,1 \cdots, D$.

Step 3: Let G denote the set of all nodes directly connected to node v_i. After n iterations of information walks, traversing G yields the information volume of node v_i:

$$H_n(v_i) = \sum \frac{H_{n-1}(v_j)}{I_0 k_j} g a_{ij} \tag{6}$$

where, a_{ij} is the connection between nodes i and j, $n = 0,1 \cdots, D$, $v_j \in A$.

Step 4: After D rounds of information propagation, the total information volume accumulated by each node in the network is computed, yielding the walk betweenness for each node. The resulting set of walk betweenness values is

$$H_G = [H(1), H(2), \cdots, H(N)] \tag{7}$$

B. Dependency Strength Based on Node Walk Betweenness

Dependent edges reflect the relationship of information exchange between subnetworks. The more frequent the exchange, the closer the dependency and the greater the mutual influence between networks. Dependency strength is commonly used to quantify the degree of such interdependence. The strength of dependency not only reflects the extent of mutual influence and control between networks but also affects the propagation speed of cascading failures within interdependent networks.

Most existing studies define dependency strength based on node degrees. However, this only captures local connectivity and ignores indirect influence, thus failing to measure the intermediary role of nodes in the global network. A few studies use betweenness centrality, assuming that information is transmitted only along shortest paths, thereby overlooking the diversity and randomness of path selection during information transmission. By simulating the random walk process, walk betweenness more accurately characterizes functional dependencies and path diversity among nodes. Accordingly, we defined a new metric for dependency strength based on walk betweenness, as follows:

$$DS(A_i, B_j) = b_i b_j + \alpha \tag{8}$$

where, $b_i = H_A(i)$ is the walk betweenness of node i in subnetwork A and $b_j = H_B(j)$ is the walk betweenness of node j in subnetwork B. The parameter α $(0 < \alpha < 1)$ is an adjustment coefficient introduced to avoid cases where edge nodes may have a walk betweenness of zero.

C. Node Operational Attribute Matching

Sensing nodes refer to combat units equipped with early warning, detection, reconnaissance, and surveillance capabilities, such as early warning radars and reconnaissance radars. Command-and-control nodes are responsible for intelligence fusion, command dissemination, information coordination, and decision support, such as command organizations and intelligence processing agencies. Firepower nodes refer to combat units capable of interception, attack, and destruction. Because these three types of nodes play different roles in the command-and-control process, their operational attributes are heterogeneous. Therefore, quantifying node operational attributes and analyzing attribute matching provides a basis for constructing functionally coordinated interlayer dependency relationships.

In this study, node operational attributes are modeled as scenario-based normalized descriptors constructed according to node roles, hierarchical characteristics, and combat-task semantics in the considered air-defense command-and-control scenario. These attributes are not intended to reproduce exact equipment specifications; instead, they provide a unified and comparable description of heterogeneous node characteristics for interlayer matching analysis. Specifically, for sensing nodes, the attribute vector mainly characterizes detection-related capability; for command-and-control nodes, it mainly characterizes communication and command-support capability; and for firepower nodes, it mainly characterizes strike-related capability.

In the upper-layer network, a node's operational capability depends not only on its intrinsic capability but also on its ability to cooperate with neighboring nodes. Therefore, the operational capability of a node in the upper-layer network is defined as

$$COM_A(i) = \sum_{i=1}^{n} \omega_i att_i + \sum_{k \in z_i} c_{ik} \quad (9)$$

where, $\sum_{i=1}^{n} \omega_i att_i$ is the weighted sum of node i ' s operational attributes, quantifying its intrinsic operational capability as an independent combat unit, and $c_{ij} = \sqrt{\sum_{z=1}^{n} (att_z^i - att_z^j)^2}$ is the cooperation factor between node i and its neighboring nodes, indicating the degree of difference between the node and its neighboring nodes in terms of attribute vectors. This factor helps compensate for individual capability gaps and reflects the cooperative capability of node i in the upper-layer network. In the simulation, the components of the attribute vector within the same node category are normalized and assigned equal weights.

In the lower-layer network, nodes mainly function as command-and-control units. Their role is not direct combat execution, but command dissemination, information transmission, and operational support. Therefore, the capability of a lower-layer node is characterized in terms of command-support capability rather than direct combat capability. Considering that the communication capability of a network node is determined by its inherent hardware characteristics and is positively correlated with its hierarchical level [18], the capability of a lower-layer command-and-control node is defined as:

$$C_COM_B(i) = \frac{CP(i) gH_i(i)}{CP_{\max} gH_{\max}} \quad (10)$$

where, $CP(i) = B(i)\log_2(1 + SNR(i))$ represents the Shannon channel capacity of node i, indicating its physical-layer information throughput capability; $H(i)$ denotes the hierarchical level of node i; both $B(i)$ and $SNR(i)$ are elements of the node's attribute vector attribute_$V(i)$.

By integrating the node operational attributes of both upper and lower networks, the interlayer node operational attribute matching is computed as follows:

$$MS_{ij} = \frac{\frac{M_{ij}}{AT_{ij}}}{\sum_{j \in B} M_{ij}/AT_{ij}} \quad (11)$$

where, $M_{ij} = \frac{1}{1+|COM_A(i) - COM_B(j)|}$ is the matching index between nodes i and j; $AT_{ij} = \frac{L_{ij}}{\max_{i,j} L_{ij}}$ is the normalized path length between nodes i and j. A larger matching value indicates that the two nodes are more compatible in terms of operational role and transmission convenience and are therefore more suitable for establishing an interlayer dependency relationship.

D. Interlayer Coupling Strategy Based on Dependency Strength and Operational Attribute Matching

In contrast to intra-layer edge strategies, interlayer edge strategies must account for both the dependency strength between nodes and their operational attribute matching. Dependency strength describes the inter-node relationship from the perspective of multipath information transmission, while attribute matching quantifies node compatibility based on operational attributes. To this end, we proposed an interlayer coupling strategy for two-layer interdependent C2Ns that integrates dependency strength and operational attribute matching. In addition, to prevent load imbalance and reduce the risk of cascading failures, an adjustment coefficient was introduced. This ensures that sensing nodes preferentially connect to upper-layer nodes in the command-and-control subnetwork, while firepower nodes preferentially connect to lower-layer nodes in the same subnetwork. The edge connection probability for the interlayer coupling strategy is calculated as

$$p_{ij} = \begin{cases} \frac{DS_{ij} MS_{ij}}{\sum_{j \in N_c} DS_{ij} MS_{ij}}, type(A_i) \in T_s \\ \frac{DS_{ij} MS_{ij} H(j)}{\sum_{j \in N_c} DS_{ij} MS_{ij} H(j)} type(A_i) \in T_f \end{cases} \quad (12)$$

where, i is a node in the upper-layer network; j is a node in the lower-layer network; N_c is the set of command-and-control nodes; $H(j)$ is the hierarchical level of node j; T_s and T_f are the sets of sensing and firepower nodes, respectively. The procedure for calculating the edge connection probability under the proposed interlayer coupling strategy is as follows (Algorithm 1):

Algorithm 1: Coupling strategy

Input: Network A, network B, attribute vectors of nodes: attribute_$V(A)$, attribute_$V(B)$

Output: Coupled network (A, B)

Step 1: Initialize nodes in network A and form queue Q_A.

Step 2: Initialize nodes in network B and form queue Q_B.

Step 3: Calculate the walk betweenness for each node using Eq. (4)-(6).

Step 4: Sequentially select from Q_A the nodes that have not yet formed coupling edges, denoted as node i. Once Q_A is fully traversed, proceed to Step 10.

Step 5: Use Eq. (8) and Eq. (11) to compute the dependency strength and attribute matching between node i and all nodes in Q_B.

Step 6: Determine the type of node i, compute its dependency probabilities with nodes in Q_B using Eq. (12), and select a dependent node j via the roulette selection method.

Step 7: Identify the neighboring nodes of i and j, denoted as Q_i and Q_j, respectively.

Step 8: Remove node j from queue Q_B. If Q_B is empty, return to Step 2.

Step 9: Treat queues Q_i and Q_j as local neighborhoods and repeat Steps 5–6. Continue this process until all Q_i pairs are traversed, then return to Step 4.

Step 10: End.

Time Complexity: $O(mn)$, where m and n are the total number of nodes in the upper and lower layers of the two-layer structure, respectively.

The above complexity analysis shows that the computational burden of the proposed method mainly arises from path-related structural evaluation and interlayer attribute matching. As the network scale increases, the structural computation becomes the dominant cost, which makes the current implementation more suitable for offline architecture optimization and periodic survivability assessment than for strict real-time operational scenarios. Nevertheless, the proposed framework remains

extensible to larger-scale command-and-control networks. Its scalability can be further improved through candidate-pair pruning, caching, or local updating of structural information, and parallel implementation of the evaluation process.

V. Simulation Analysis

To validate the effectiveness of the proposed strategy, we conducted simulation experiments, covering parameter analysis, interlayer coupling strategy analysis, and network model analysis. Based on an air defense command and control system, the initial parameter values were as follows: the total number of nodes is 580, including 160 command and control nodes (n_c = 160), 150 sensing nodes (n_s = 150), and 270 firepower nodes (n_f = 270); the command span was set as k = 3 and the command hierarchy as h = 4. For all simulations, the components of each node's attribute vector were normalized and assigned equal weights within the same node category to control the number of free parameters and ensure a fair comparison among different coupling strategies.

Two attack strategies, namely deliberate attack and random attack, were adopted. Unless otherwise specified, each reported result was obtained from 200 independent simulation runs. For each metric, the mean value across runs is reported, and the shaded region in the corresponding figures represents the 95% confidence interval. In this way, the robustness of the observed performance differences can be evaluated not only in terms of average behavior but also in terms of statistical variability across repeated simulations.

A. Analysis of Dependency Strength Adjustment Parameter α

To determine a suitable value for the dependency strength adjustment parameter α, we conducted simulation experiments to assess network performance as α varies. Given the large scale of the network (580 nodes in total), a significant number of nodes must be removed to clearly observe changes in node survival rate. Fig. 2 shows the variation in node survival rate with α after removing 100–150 nodes under deliberate attack. Fig. 3 shows the same under random attack with 300–350 nodes removed. From Fig. 2 and 3, clearly, the node survival rate peaks at α = 0.3. Therefore, this value is used in all subsequent experiments.

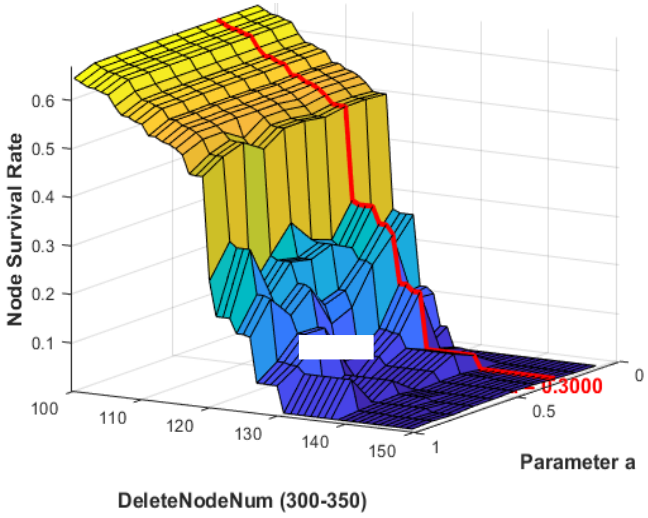


Fig. 2. Node survival rates under different α values (deliberate attack).

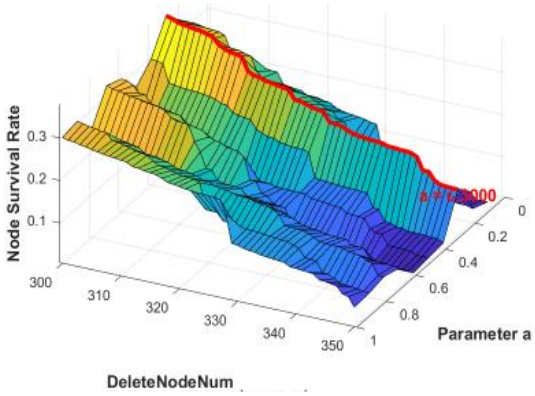


Fig. 3. Node survival rates under different α values (random attack).

B. Analysis of Interlayer Coupling Strategies

To validate the rationality of the proposed interlayer coupling strategy, we conducted comparative simulations focusing on different dependency strengths, node operational attribute matching levels, and interlayer coupling connection strategies.

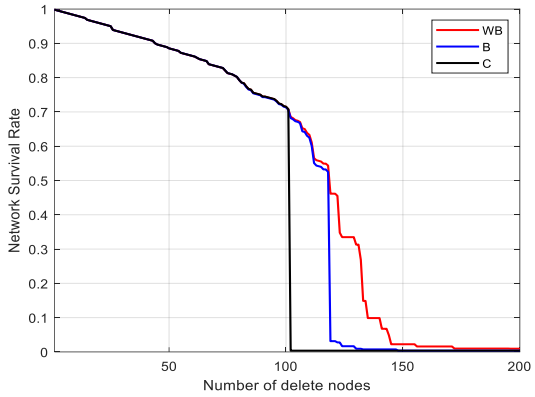


Fig. 4. Node survival rates under different dependency strengths (deliberate attack).

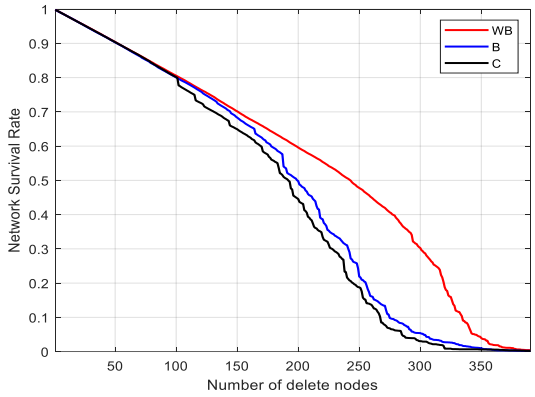


Fig. 5. Node survival rates under different dependency strengths (random attack).

1) Comparative analysis of different dependency strengths: Three types of dependency strengths were selected for comparison: walk betweenness (WB), betweenness centrality (B), and closeness centrality (C). Fig. 4 and Fig. 5 show the node survival rates under deliberate and random attacks,

respectively, for these three strategies. Under deliberate attack (Fig. 4), the node survival rates of all three methods were initially similar. However, after removing 100 nodes, an inflection point appeared, indicating the onset of large-scale cascading failures. For networks constructed using betweenness and closeness centrality, node survival rapidly dropped to zero. In contrast, the dependency strength based on WB proposed in this study delayed cascading failures. Its survival rates did not drop to zero until 150 nodes were removed. Under random attack (Fig. 5), the proposed method clearly outperformed the other two. This advantage is attributed to the use of WB, which accounts for multipath information transmission and dynamic network behavior, thereby mitigating abrupt performance degradation.

2) Comparative analysis of different node matching degrees: Three scenarios were compared: operational attribute matching (AM) proposed in this study, matching based on the node's hierarchical level (LM), and random matching (RM). Fig. 6 and 7 illustrate the node survival rates under deliberate and random attacks, respectively, for these three matching strategies. Under both attack strategies, the proposed matching degree achieved the highest node survival rate and the best network survivability. This is because considering node matching degree helps avoid unbalanced dependency edges formed between nodes with significantly different combat capabilities, thereby suppressing cascading failures.

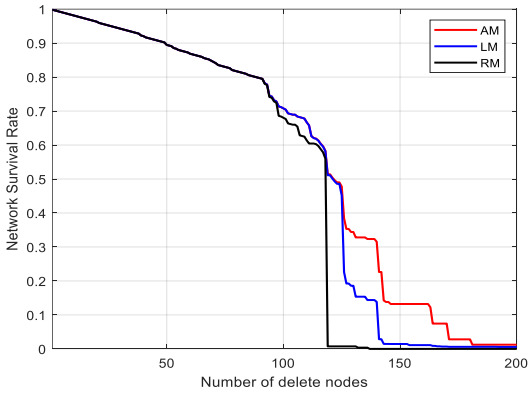


Fig. 6. Node survival rates under different node matching degrees (deliberate attack).

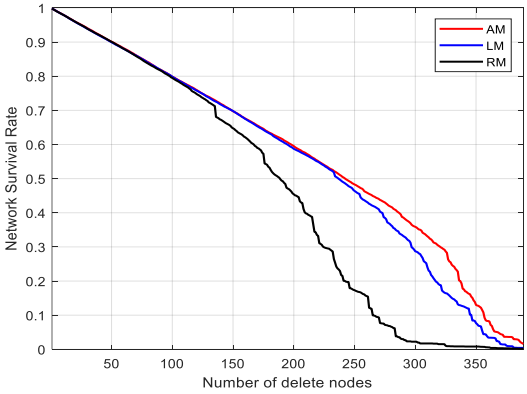


Fig. 7. Node survival rates under different node matching degrees (random attack).

3) Comparative analysis of different interlayer coupling strategies

a) Node survival rate analysis: Eight interlayer coupling strategies were evaluated: the proposed strategy (PRO), the strategy from Ref. [16] (DB), random (RL), neighbor NPC, degree-assortative coupling (BAL), degree-disassortative coupling (BDL), betweenness-assortative coupling (DAL), and betweenness-disassortative coupling (DDL). Fig. 8 and 9 show the node survival rates under deliberate and random attacks, respectively. As can be seen in the figures, under both attack strategies, the proposed PRO strategy achieved the highest node survival rate and better network survivability. This is because WB captures the global influence of a node under multipath and random walk conditions, rather than relying solely on the shortest path. The importance of key nodes is distributed across more paths instead of being concentrated in a few dominant nodes. Therefore, even after partial node failures, the network can still maintain functionality through non-shortest paths, thereby delaying the sharp decline in node survival rate.

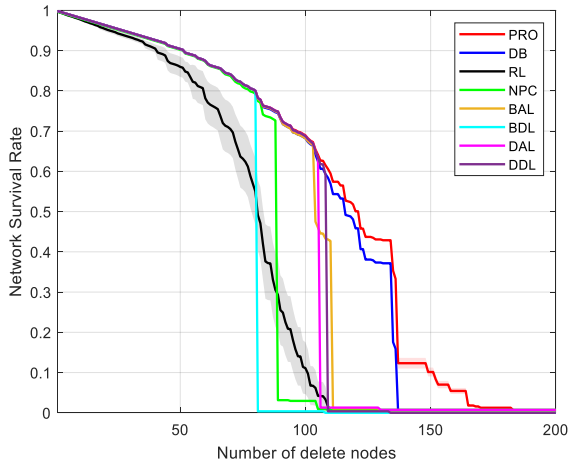


Fig. 8. Node survival rates under different interlayer coupling strategies (deliberate attack).

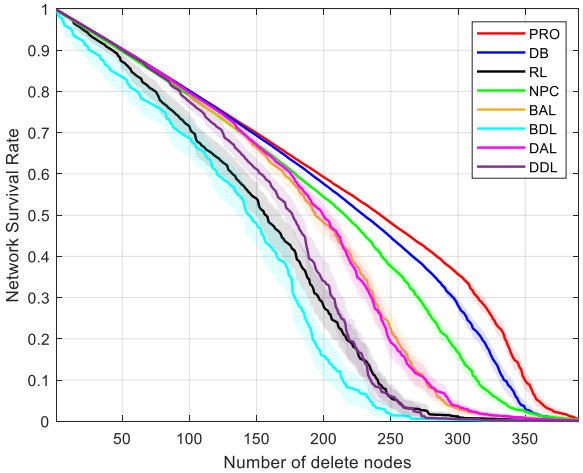


Fig. 9. Node survival rates under different interlayer coupling strategies (random attack).

b) OODA-Oriented functional chain retention analysis: To further evaluate whether the proposed strategy can preserve the functional stability of the command-and-control process, an OODA-oriented functional chain retention ratio is introduced. This metric is used to measure the extent to which the network can still maintain effective sensing–command–firepower coordination after node failures and cascading effects.

Specifically, an effective OODA-oriented functional chain is considered to exist if a surviving sensing node can still coordinate with a surviving command-and-control node, and the latter can further support a surviving firepower node through the remaining intra-layer and interlayer connections. Then, the OODA-oriented functional chain retention ratio is defined as the ratio between the number of effective functional chains after attack and that in the initial network.

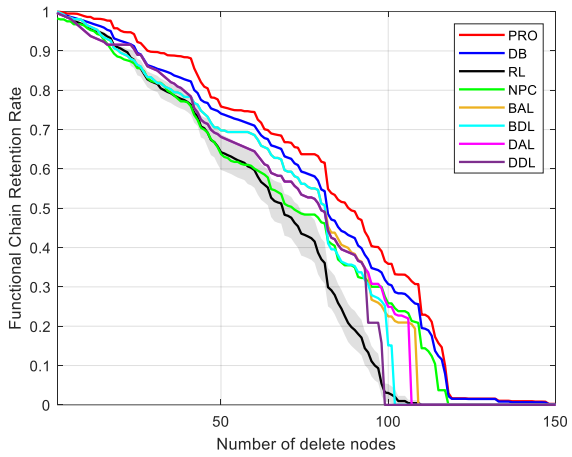


Fig. 10. OODA-oriented functional chain retention ratios under different coupling strategies (deliberate attack).

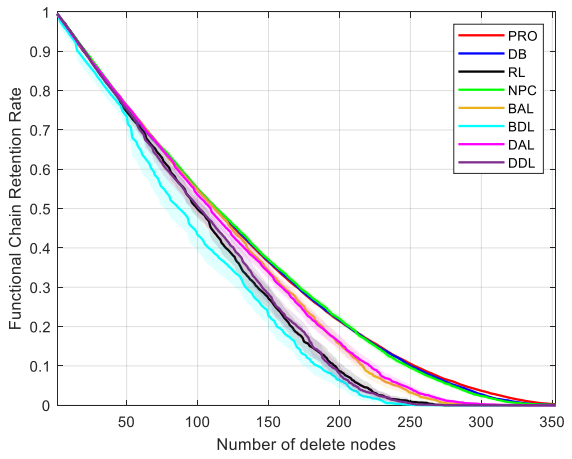


Fig. 11. OODA-oriented functional chain retention ratios under different coupling strategies with 95% confidence intervals (random attack).

Fig. 10 and Fig. 11 show the OODA-oriented functional chain retention ratios under deliberate and random attacks, respectively. As can be seen from the figures, under both attack strategies, the PRO strategy maintained the highest functional chain retention ratio and exhibited the slowest degradation trend among all compared methods. This indicates that the proposed strategy is more effective in preserving the sensing–command–firepower coordination required by the command-and-control process, rather than merely retaining a larger number of surviving nodes.

Compared with the baseline strategies, the advantage of PRO in this metric can be attributed to two aspects. First, the walk-betweenness-based dependency strength better captures multipath information transmission characteristics, thereby reducing the vulnerability caused by over-reliance on shortest-path-dominated key nodes. Second, the operational attribute matching mechanism reduces functionally mismatched interlayer couplings, which helps preserve semantically compatible sensing–command–firepower chains under attacks. Therefore, the superiority of PRO is reflected not only in structural robustness but also in functional robustness related to OODA-oriented command-and-control stability.

c) Combat link efficiency analysis: The same eight strategies were used to compare combat link efficiency. Fig. 12 and Fig. 13 show the degradation rate of combat link efficiency under deliberate and random attacks, respectively. As seen in the figures, under both attack strategies, the PRO strategy displayed the best combat link efficiency, with lower degradation rates of combat link efficiency than other models. This is because the proposed strategy considers node operational attributes, which reduces functional mismatches between interlayer-coupled nodes and improves overall network operational effectiveness. These results further validate the rationality of the proposed model.

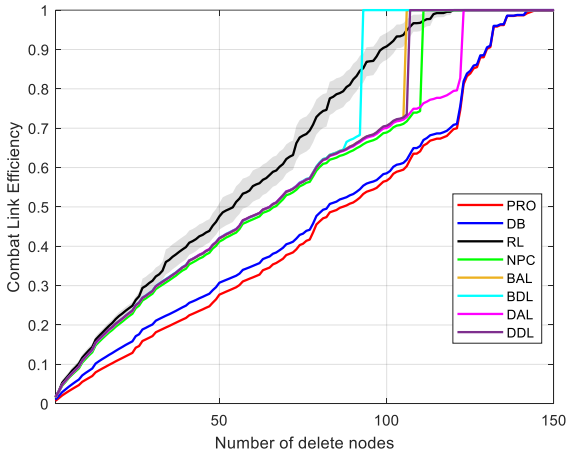


Fig. 12. Combat link efficiency degradation rate under different coupling strategies (deliberate attack).

d) Comprehensive analysis: Tables II and III summarize the average node survival rate, OODA-oriented functional chain retention ratio, and combat link efficiency degradation rate of different interlayer coupling strategies under deliberate and random attacks, respectively. All reported values are the mean results over 200 independent simulation runs, while the corresponding line figures present the 95% confidence intervals as shaded bands. Degree-disassortative coupling and betweenness-disassortative coupling strategies were excluded because of their significantly inferior overall performance.

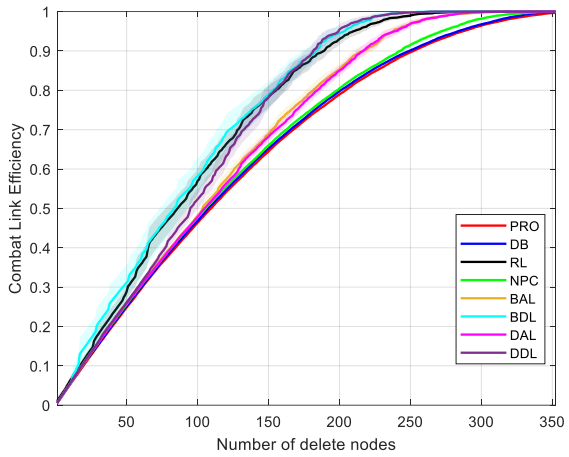


Fig. 13. Combat link efficiency degradation rate under different coupling strategies (random attack).

The results show that the PRO strategy achieves the best overall performance under both attack modes. Under deliberate attack, PRO attains the highest node survival rate and the lowest combat link efficiency degradation rate, while maintaining a competitive functional chain retention ratio close to that of the best-performing baseline. Under random attack, PRO performs best among all compared strategies in all three metrics.

Compared with the baseline random strategy (RL), PRO improves node survival rate by 24.82% and 50.00% under deliberate and random attacks, respectively. Meanwhile, the OODA-oriented functional chain retention ratio is improved by 13.38% and 19.40%, and the combat link efficiency degradation rate is reduced by 6.37% and 3.53%, respectively. These results indicate that the superiority of PRO is consistently reflected in structural robustness, functional chain preservation, and performance maintenance.

TABLE II. COMPARISON OF NETWORK PERFORMANCE UNDER DIFFERENT COUPLING STRATEGIES (DELIBERATE ATTACK)

	Average node survival rate (%)	Average functional chain retention ratio (%)	Average combat link efficiency degradation rate (%)
PRO	58.79	19.58	71.23
DB	56.39	20.76	71.67
NPC	54.61	17.58	73.26
BAL	51.24	18.15	74.21
DAL	48.12	17.90	75.17
RL	47.10	17.27	76.08

The advantage of PRO mainly stems from two aspects. First, walk betweenness characterizes node importance from the perspective of multipath information transmission, thereby alleviating the excessive reliance on shortest-path-dominated structural indicators in conventional coupling methods. Second, the operational attribute matching mechanism improves functional compatibility between cross-layer nodes, which helps preserve sensing–command–firepower coordination and suppress the propagation of cascading failures.

TABLE III. COMPARISON OF NETWORK PERFORMANCE UNDER DIFFERENT COUPLING STRATEGIES (RANDOM ATTACK)

	Average Node survival rate (%)	Average functional chain retention ratio (%)	Average combat link efficiency degradation rate (%)
PRO	56.55	32.07	87.10
DB	53.74	31.91	87.22
NPC	49.98	32.00	89.05
BAL	44.79	29.58	89.36
DAL	43.85	29.54	89.55
RL	37.70	26.86	90.29

C. Network Model Analysis

To further evaluate the advantages and disadvantages of the proposed edge connection strategy, we constructed six network models (Table IV). Among them, M1, M2, M4, and M5 adopt the attribute synergy strategy for constructing the upper-layer subnetworks (sensing subnetwork and firepower subnetwork), while M3 and M6 apply the scale-free edge connection strategy. In the detect–strike integration phase, M1 and M4 employ the node efficiency prioritization strategy, whereas M2, M3, M5, and M6 adopt the degree prioritization strategy. All models use an improved tree-like network model for the intra-layer lower-layer subnetwork (command and control subnetwork). Based on these models, two categories of experiments were designed: (i) To compare the performance of the proposed strategy with that of the method in Ref. [16] under the same subnet construction configurations, with comparison groups including M1 vs. M4, M2 vs. M5, and M3 vs. M6; (ii) To evaluate the applicability of the proposed strategy under different subnet construction models, with comparison groups including M1, M2, and M3.

TABLE IV. CONSTRUCTION STRATEGIES FOR NETWORK MODELS

Model	Intra-layer Upper Network Linking Strategy		Intra-layer Lower Network Linking Strategy	Inter-layer Coupling Strategy
	Perception–Firepower Subnet Linking Strategy	Integrated Reconnaissance–Strike Operation Mode Linking Strategy		
M1 (Pro Model)	Attribute-prioritized	Node-efficiency-prioritized	Improved tree topology	Proposed method
M2	Attribute-prioritized	Degree-prioritized	Improved tree topology	Proposed method
M3	Scale-free	Degree-prioritized	Improved tree topology	Proposed method
M4	Attribute-prioritized	Node-efficiency-prioritized	Improved tree topology	Method of Ref. [16]
M5	Attribute-prioritized	Degree-prioritized	Improved tree topology	Method of Ref. [16]
M6	Scale-free	Degree-prioritized	Improved tree topology	Method of Ref. [16]

1) Node survival rate analysis: Under random attack (Fig. 15), all models had survival rate curves that were nearly

linear with similar slopes in the early stages. In the later stages, under the same intra-layer connection strategy, the proposed strategy slightly outperformed the method in Ref. [16], with only minor differences between the curves. This suggests that during random node removal, even with different coupling strategies, each model can rely on the inherent scale-free redundancy of the network to maintain connectivity in the early phase.

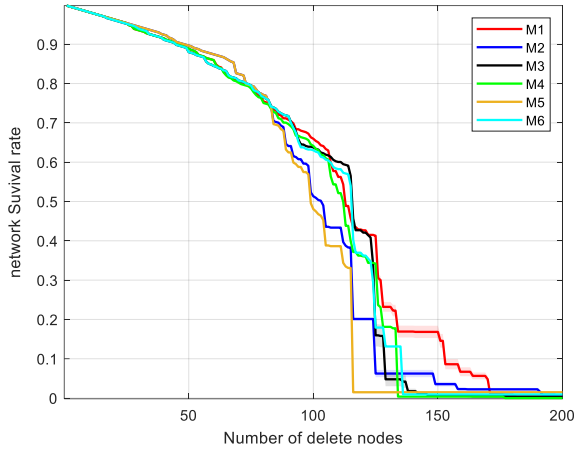


Fig. 14. Node survival rates of different network models (deliberate attack).

Under deliberate attack (Fig. 14), inflection points appeared in the node survival rate curves during the mid-to-late phases. A steep "step-wise decline" occurred after every few dozen node removals, corresponding to the cascading failure effect in interdependent networks. Under deliberate attack, M3 and M6 were the first to collapse. Their survival rates dropped sharply when 120 nodes were removed. M2 and M5 followed, collapsing when 150 nodes were deleted. M4 failed around 160 node removals. M1 had the highest stability: it survived until 180 nodes were removed. Thus, the proposed strategy performed best under conditions of attribute synergy and node efficiency prioritization, but was slightly less effective in scale-free configurations. As the proposed strategy introduces operational AM, it effectively optimizes node selection based on node operational attributes. However, when subnets are constructed using scale-free connections, this introduces interference to the strategy. In contrast, scale-free subnetworks inherently reduce overall network survivability. Comparisons among M1 vs. M4, M2 vs. M5, and M3 vs. M6 demonstrated that the proposed strategy consistently outperformed the method in Ref. [16] under various subnet modeling conditions, highlighting its robustness and general applicability.

2) OODA-Oriented functional chain retention analysis: To further examine whether the proposed strategy can maintain functional stability under different network structures, the OODA-oriented functional chain retention ratios of different network models were compared. Fig. 16 and Fig. 17 show the corresponding results under deliberate and random attacks, respectively.

As can be seen from the figures, although the overall degradation patterns vary across network models, the PRO strategy consistently maintains a higher OODA-oriented functional chain retention ratio than the compared methods in most cases. This indicates that the proposed strategy is able to preserve sensing–command–firepower coordination more effectively under different structural configurations, rather than exhibiting advantages only in a specific network instance.

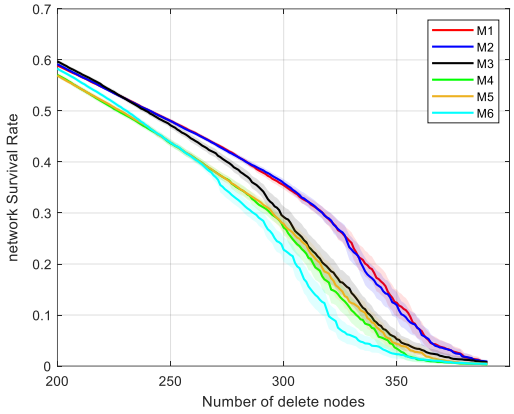


Fig. 15. Node survival rates of different network models (random attack).

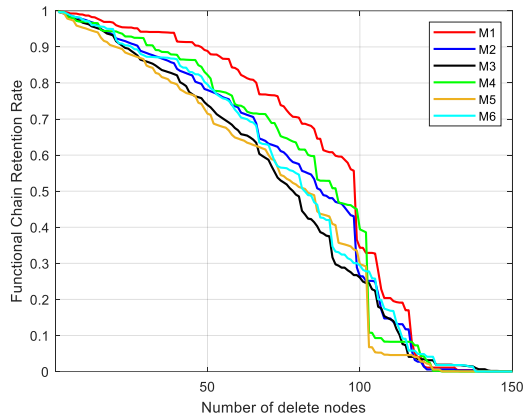


Fig. 16. OODA-oriented functional chain retention ratios under different network models (deliberate attack).

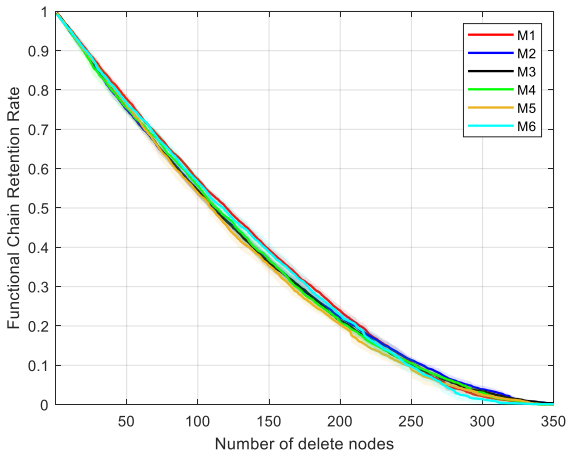


Fig. 17. OODA-oriented functional chain retention ratios under different network models (random attack).

3) Combat link efficiency analysis: Under both random and deliberate attacks (Fig. 18 and Fig. 19), M1 consistently maintained the lowest degradation rate of combat link efficiency. As the number of removed nodes increased, the proposed model demonstrated a clear advantage in terms of the degradation rate of combat link efficiency, compared to the other three models. This is attributed to the fact that the proposed strategy more effectively selects nodes with matched combat capabilities across the two subnetworks, particularly within the attribute-synergy subnetwork, thereby reducing combat capability imbalances between nodes and enhancing overall network robustness.

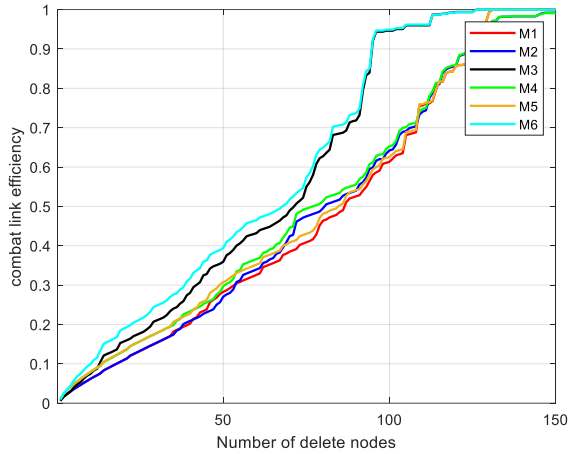


Fig. 18. Combat link efficiency degradation rate of different network models (deliberate attack).

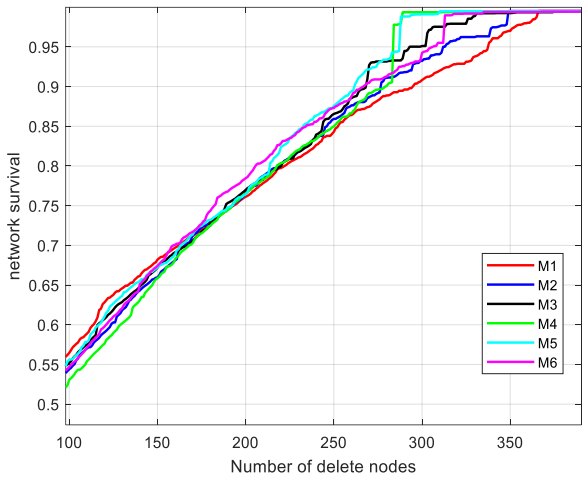


Fig. 19. Combat link efficiency degradation rate of different network models (random attack).

VI. Discussion

The simulation results demonstrate that the advantage of PRO is not limited to a single metric or a specific attack mode. Compared with topology-driven coupling strategies, PRO achieves better overall performance in node survival rate, OODA-oriented functional chain retention ratio, and combat link efficiency degradation rate. This indicates that the proposed strategy improves both the structural survivability of interdependent C2Ns and the functional stability of sensing–command–firepower coordination under cascading failures.

The performance improvement mainly comes from the joint consideration of multipath structural importance and operational compatibility. The walk-betweenness-based dependency strength captures the intermediary role of nodes from the perspective of multipath information transmission, rather than relying only on local degree or shortest-path-based indicators. Therefore, it helps avoid excessive concentration of dependency risks on a small number of structurally dominant nodes. Meanwhile, the operational attribute matching mechanism improves the functional rationality of interlayer dependency edges. Since sensing, command-and-control, and firepower nodes perform different roles in the OODA process, functionally compatible cross-layer coupling is important for maintaining effective operational chains after node failures.

From a system-engineering perspective, the results suggest that interlayer dependency design in C2Ns should jointly consider structural importance and role-related operational compatibility. The proposed strategy is particularly suitable for C2N architecture planning, survivability-oriented network design, and periodic robustness assessment before deployment or reconfiguration. Although the computation of walk betweenness and attribute matching introduces additional cost, the framework can be further accelerated through candidate-pair pruning, caching or local updating of structural indicators, and parallel implementation of the matching process.

VII. Future Directions

Future work will further extend the proposed framework from three aspects. First, more realistic C2N data and equipment-level operational parameters can be incorporated to improve the fidelity of node attribute modeling, including sensing, communication, command-support, and firepower-related characteristics. Second, the current interlayer coupling design can be extended to dynamic and adaptive scenarios, where communication quality, node status, task priorities, and command relationships may change over time. Third, the computational efficiency of the proposed framework can be further improved for larger-scale C2Ns through candidate-pair pruning, incremental updating of structural indicators, distributed computation, and parallel matching algorithms. These extensions would further enhance the scalability and applicability of PRO in larger-scale and more dynamic command-and-control network scenarios.

VIII. Conclusion

This study proposed PRO, an interlayer coupling strategy for two-layer interdependent C2Ns based on node walk betweenness and operational attributes. The proposed strategy integrates walk-betweenness-based dependency strength with operational attribute matching, thereby supporting survivability-oriented interlayer dependency design from a joint structural–functional perspective.

Simulation experiments on an air-defense command-and-control scenario were conducted under deliberate and random attacks. The results show that PRO achieves better overall performance than the compared baseline strategies in terms of node survival rate, OODA-oriented functional chain retention

ratio, and combat link efficiency degradation rate. These findings indicate that the proposed strategy can effectively mitigate cascading failure risks while preserving sensing–command–firepower coordination.

The main contribution of this work lies in providing a structural–functional coupling framework for interdependent C2Ns. Unlike topology-only coupling strategies, PRO jointly considers multipath information transmission and operational compatibility, which provides a more mission-oriented basis for interlayer dependency design in command-and-control networks.

REFERENCES

[1] Y. Li, Y. Sun, M. Xie, Y. Sun, and X. Zhou, "Nested OODA: A C2 model based on fractal theory," in *Proc. Int. Conf. Networking, Sensing and Control (ICNSC)*, Hangzhou, China, 2024, pp. 1–6.

[2] R. Yin, L. Li, Y. Wang, Z. Hao, and L. Zhang, "Robustness of networks with dependence clusters against hybrid cascading failure induced by fluctuating load," *Physica A: Statistical Mechanics and its Applications*, vol. 660, Art. no. 130367, 2025.

[3] L. Yang, Y. Dang, G. Guo, and J. Jiang, "Robustness analysis of interdependent network accounting for failure probability and coupling patterns," *Chaos*, vol. 34, no. 6, Art. no. 063134, 2024.

[4] G. Dong, N. Sun, M. Yan, F. Wang, and R. Lambiotte, "Robustness of coupled networks with multiple support from functional components at different scales," *Chaos*, vol. 34, no. 4, Art. no. 043122, 2024.

[5] X. Xu and X. Fu, "Analysis on cascading failures of directed-undirected interdependent networks with different coupling patterns," *Entropy*, vol. 25, no. 3, p. 471, 2023.

[6] L. Liu, Y. Yin, Z. Zhang, and Y. K. Malaiya, "Redundant design in interdependent networks," *PLoS ONE*, vol. 11, no. 10, Art. no. e0164777, 2016.

[7] Z. Chen and W. Du, "Cascading failure of interdependent networks with different coupling preference under targeted attack," *Chaos, Solitons & Fractals*, vol. 80, pp. 7–21, 2015.

[8] Y. Hao, L. Jia, and Y. Wang, "Cascading failures in interdependent scale-free networks of different coupling preferences with the harmonic closeness," *EPL (Europhysics Letters)*, vol. 127, no. 3, Art. no. 38003, 2019.

[9] X. Wang, W. Zhou, R. Li, J. Cao, and X. Lin, "Improving robustness of interdependent networks by a new coupling strategy," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 1075–1080, 2018.

[10] W. Zang, X. Ji, S. Liu, and H. Li, "An interdependent network coupling strategy based on overlapping link structure against targeted attack," *International Journal of Modern Physics C*, vol. 32, no. 8, Art. no. 2150101, 2021.

[11] C. Li, X. Xu, and Y. Zhang, "Robustness analysis of edge-coupled interdependent networks under different coupling modes," *Physica A: Statistical Mechanics and its Applications*, vol. 613, Art. no. 128613, 2023.

[12] I.-C. Lin, O. Yagan, and C. Joe-Wong, "Evaluating the optimality of dynamic coupling strategies in interdependent network systems," in *ICC 2023 - IEEE Int. Conf. Communications*, Rome, Italy, 2023, pp. 1474–1480.

[13] Y. Gao, H. He, J. Liu, and S. Chen, "Percolation behaviors of partially edge-coupled interdependent networks," *Physics Letters A*, vol. 431, Art. no. 127919, 2022.

[14] X. Ji, B. Wang, and D. Liu, "Improving interdependent networks robustness by adding connectivity links," *Physica A: Statistical Mechanics and its Applications*, vol. 444, pp. 9–19, 2016.

[15] Y. Wang, S. Chen, C. Pan, and B. Chen, "Measure of invulnerability for command and control network based on mission link," *Information Sciences*, vol. 426, pp. 148–159, 2018.

[16] B. Chen, G. Pang, Z. Xiang, X. Gao, Y. Chen, and S. Chen, "Modeling dual-layer interdependent command and control networks for integrated reconnaissance-strike and OODA-loop capabilities," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 6, pp. 5744–5759, 2024.

[17] B. Chen, H. Tao, X. Jiang, Y. Chen, and X. Gao, "Dynamic hypernetwork-based evolutionary model of command-and-control network," *Simulation Modelling Practice and Theory*, vol. 126, Art. no. 102759, 2023.

[18] C. Pan, W. Xiong, S. Qiu, Y. Wang, and B. Chen, "Modeling method of command and control network based on multi-attribute weighting," *Modern Defence Technology*, vol. 46, no. 4, pp. 60–66, 2018.

[19] M. F. Ling, T. Moon, and E. Kruzins, "Proposed network centric warfare metrics: From connectivity to the OODA cycle," *Military Operations Research*, vol. 10, no. 1, pp. 5–13, 2005.

[20] B. Zhao, M. Wang, Q. Zhang, W. Lin, and D. Zhou, "Research on the robustness of command and control networks under cascading failures," *Electronics*, vol. 13, no. 20, Art. no. 3992, 2024.

Utilizing Structured Equation Modeling and Machine Learning in Investigating Digital Competency in Public School Teachers in Bukidnon, Philippines

Nathalie Joy G. Casildo
Information Technology Dept.-College of Information Sciences and Computing,
Central Mindanao University, Bukidnon, Philippines

***Abstract*—The rapid transition to digital education in the Philippines, accelerated by the COVID-19 pandemic, has highlighted significant integration challenges for public school teachers in rural provinces like Bukidnon. While digital proficiency is essential, existing studies often rely on either purely descriptive analytics or standalone machine learning models, which frequently fail to validate the complex, latent relationships between competency factors in terms of Digital Competency. To address these gaps, this research employs a two-phased hybrid analytical architecture. Phase I utilizes Structural Equation Modeling (SEM) to confirm the factor structure and establish the causal pathways of digital competency, ensuring that the theoretical framework is psychometrically sound. Phase II transitions these validated constructs into an optimized Machine Learning (ML) pipeline, incorporating SMOTE-ENN resampling to handle imbalanced regional data. Results from 1,275 participants demonstrate that "Professional Engagement" acts as the foundational engine of the digital competency system, while "Digital Pedagogy in Teaching" emerges as the most critical predictive determinant of teacher proficiency. The Random Forest algorithm achieved a high predictive accuracy of 89% and a Macro F1-Score of 85%, significantly outperforming traditional models. These findings indicate that the digital divide in this context is a pedagogical, rather than purely technical, bottleneck. The study provides a blueprint for the Department of Education to move from descriptive reporting toward Predictive Diagnostic Systems that can facilitate targeted, data-driven interventions.**

Keywords*—Structural Equation Modeling (SEM); Machine Learning (ML); digital competence; DigCompEdu; teacher proficiency; SMOTE-ENN; predictive analytics; educational data mining

I. Introduction

Digital transformation in the Philippines was historically gradual, only accelerating rapidly out of necessity during the COVID-19 pandemic. This shift made remote work and digital education the primary modes of instruction. In response, national agencies like the Department of Education (DepEd), specifically in Region 10, mandated the full integration of Information and Communication Technology (ICT) in primary instruction. However, despite these mandates, public school teachers in inland provinces like Bukidnon continue to face significant integration challenges. Infrastructure remains a challenge; many schools still lack stable signals, and even the existence of satellite technologies is often insufficient. Beyond hardware, there is a persistent lack of awareness regarding data security and digital literacy, which suggests that infrastructure alone cannot solve the underlying threat to digital readiness.

While the Philippines has adopted various ICT frameworks, the actual digital proficiency of teachers in these rural settings remains under-explored through advanced analytics. In the Philippine educational context, most studies rely on either pure descriptive analytics or standalone Machine Learning (ML) and Educational Data Mining (EDM) [1]. While a few researchers have attempted to combine Structural Equation Modeling (SEM) and ML [2], these efforts are often hampered by limited datasets or a narrow focus on general digital literacy. Consequently, existing methods frequently fail to validate the complex, latent (hidden) relationships between competency factors, fail to accurately predict future teacher performance, and struggle to handle skewed data—where the majority of teachers might fall into a single proficiency category.

To address these gaps, this research employs a two-phased hybrid architecture. Phase I utilizes Structural Equation Modeling (SEM) to confirm the factor structure and establish the causal pathways of digital competency. Phase II then transitions these validated constructs into Machine Learning (ML) algorithms, such as Random Forest and Support Vector Machines, to build a predictive model optimized for regional educational data.

Unlike traditional studies that rely on either simple statistics or "black-box" machine learning, this study proposes a hybrid SEM-ML architecture. Using SEM as a rigorous validation layer for feature engineering, the research ensures that the inputs for the machine learning model are not just statistically significant but are theoretically grounded and psychometrically sound.

II. Related Work

A. Computational Modeling of Digital Competence

Digital competence encompasses an integration of technical, cognitive, and socio-emotional proficiencies [3]. Assessing these skills involves using established frameworks [4] paired with computational modeling [5, 6]. Applying mathematical approaches, these models provide a scientific basis for interpreting the status of an individual's digital competence [7]. However, the focus of these frameworks must move beyond technical skills toward pedagogical intervention [8].

B. Digital Competence in DepEd Teachers

Existing research affirms the value of digital literacy (DL) and digital competence (DC) for Filipino teachers, enabling them to engage with new technologies critically within digital pedagogy [9]. Enhancing these competencies can drive higher work engagement, ultimately leading to improved job performance and satisfaction, particularly in remote work environments [10]. In terms of regional applications, [4] and [11] provide significant contributions. Specifically, [4] explores frameworks in Fig. 1 relevant at national, international, and regional scales, such as the framework developed by the UNESCO Asia and Pacific Regional Office. Furthermore, the work in [11] serves as a valuable benchmark for the Bukidnon context; it analyzes digital competence frameworks used in Ibero-America through 2022 to establish a conceptual foundation for assessment tools in regional educational settings.

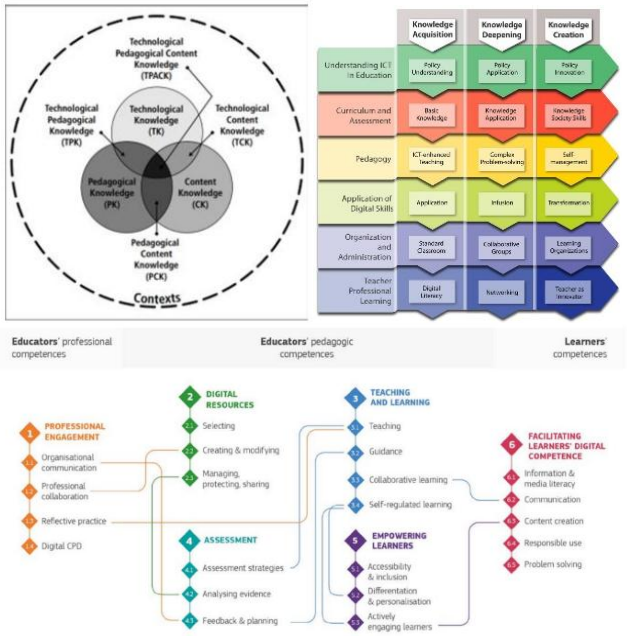


Fig. 1. ICT and competency frameworks [4].

C. Structural Equation Modeling (SEM) as a Validation Layer

Structural Equation Modeling (SEM) serves as a robust validation layer in contemporary research [12–14]. Since digital competence is a latent construct that eludes direct measurement, SEM utilizes Confirmatory Factor Analysis (CFA) to verify its underlying dimensions [15]. A practical application of this is seen in the validation of the TDC-S instrument [16]. In this study, Fig. 2 presents the CFA output for measurement model 3, reporting $X^2 = 1{,}480.693$, ($df = 584, p = 0.00$) and an X^2/df ratio of 2.535. Further evaluation of the model's performance includes an RMSEA of 0.075 and a CFI of 0.897. Following the benchmarks established in recent literature, [17] provides an interpretation of these indices to determine the scientific status of the model's fit. A study establishes a robust theoretical framework for confirming theoretical constructs and quantifying latent relationships, which serves as the foundation for subsequent predictive modeling [18].

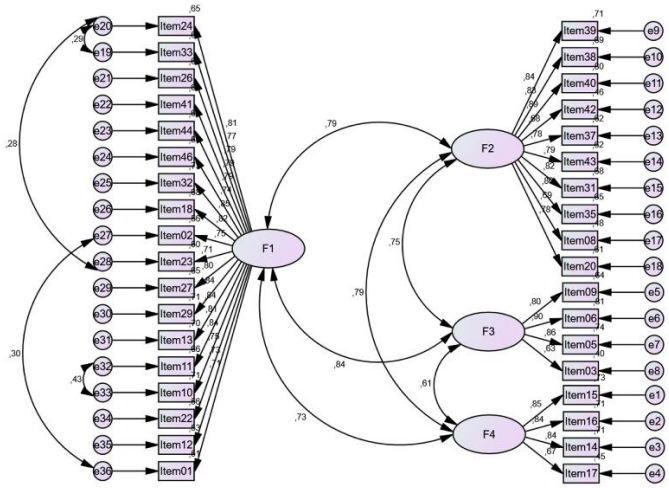


Fig. 2. CFA output of measurement model 3 [14].

D. Machine Learning and Educational Data Mining

Recent research has explored the application of machine learning (ML) to predict teachers' digital competency [19–21]. Within the field of Educational Data Mining (EDM), the most frequently utilized algorithms include Logistic Regression, Random Forest, and Decision Trees [19, 22, 23]. Furthermore, several studies have introduced hybrid methodologies that integrate Structural Equation Modeling (SEM) with Machine Learning [24, 25]. Specifically, [24] demonstrates that combining Bayesian SEM with ML achieves superior levels of predictive accuracy and robustness compared to traditional SEM or standalone ML techniques. Similarly, [25] bridges the gap between causal explanation and predictive accuracy by employing SEM to identify direct and indirect pathways, while using ML methods to reveal marginal contributions and nonlinear transition characteristics. These studies collectively indicate that the integration of SEM and ML offers a more effective and comprehensive analytical approach.

E. Hybrid SEM-ML Architectures in the Global Research Landscape

The integration of Machine Learning (ML) into Structural Equation Modeling (SEM) frameworks has emerged as a robust analytical paradigm in recent educational research. This hybrid approach allows for a fusion of explanation and prediction, addressing the limitations of theory-driven models in highly complex environments. Brandmaier and Jacobucci [26] emphasize that integrating ML techniques into SEM renders theory development more structured, efficient, and robust, particularly when dealing with high-dimensional data that traditional path analysis may struggle to parse.

Recent international studies have successfully operationalized this synergy across various educational domains. For example, Zhang et al. [27] applied a hybrid approach to predict student engagement in physical education, utilizing nine ML algorithms—with Random Forest yielding the highest predictive accuracy—to rank the influence of various digital tools. By concurrently using SEM to map causal pathways, the authors provided evidence on how technology usage indirectly impacts engagement through psychological mechanisms such as perceived usefulness and self-efficacy.

Furthermore, this hybrid architecture is instrumental in identifying the non-linear realities of digital transformation. In a large-scale study of 650 participants, a recent investigation [28] demonstrated that structural models can sometimes reveal statistically insignificant relationships (e.g., between technological infrastructure and actual digital adoption) that might be overlooked in simpler frameworks. This suggests that digital transformation is not a deterministic outcome of resource allocation or top-down mandates, but rather a complex, multi-faceted process. Positioning our study within this global methodological landscape, leveraging these advanced techniques to provide a nuanced, data-driven diagnostic of digital competency among public school teachers in Bukidnon, Philippines.

F. Resampling Techniques for Imbalanced Regional Data

Data imbalance is a prevalent challenge in educational datasets, often leading to biased models, untrustworthy outcomes, and degraded predictive performance [28–30]. To mitigate these issues, researchers frequently employ data-level interventions such as the Synthetic Minority Oversampling Technique (SMOTE). For instance, [29] implemented SMOTE-NC (for nominal and categorical data) to synthesize minority cases in high school dropout predictions, directly addressing class distribution gaps. Unlike Random Oversampling (ROS), which can lead to overfitting through sample duplication, SMOTE generates new synthetic instances by interpolating between a minority sample and its *k*-nearest neighbors [31]. Beyond oversampling, the literature suggests that various sampling strategies are essential for optimizing performance metrics [32–34]. Notably, [35] found that Random Undersampling (RUS) and hybrid techniques significantly improved Recall and F1-scores, demonstrating that prioritizing the detection of minority class instances is often more critical than maximizing overall accuracy.

III. Methodology

Fig. 3 shows the Knowledge Discovery and Predictive Analytics process. The activity started from research instrument design, where the conceptualization and schema design of the research instrument were involved. The questionnaire was adapted from the European Framework for Digital Competence in Education (DigCompEdu) to ensure construct validity. Once the framework is established, proceed to the Data Acquisition of Large-scale Survey Collection from the teachers of Bukidnon.

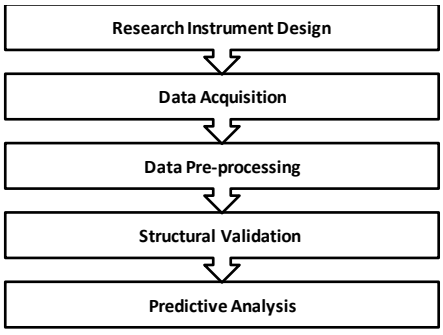


Fig. 3. Knowledge discovery and predictive analytics process.

The third phase is the Data Pre-Processing involving Data Cleaning & Feature Engineering. Proceeded with structural validation, in which Structural Equation Modeling was used using SPSS AMOS. The last step is the Predictive Analysis using logistic Regression, where accuracy calculation and Predictive classification of digital divide levels were conducted.

A. Dataset Acquisition and Characteristics

The dataset for this study was acquired through a large-scale survey involving public school teachers within the Division of Bukidnon, Philippines. Data collection was conducted between October 2022 and March 2023. A total of 1,275 valid responses were obtained, as shown in Table I: Summary of Dataset Characters. Responses represent a diverse cross-section of grade levels, school types, and geographic locations. The primary instrument was adapted from the European Framework for Digital Competence in Education (DigCompEdu), consisting of 22 items measured on a five-point Likert scale. In order to ensure the consistency of the multi-dimensional constructs, Cronbach's Alpha was calculated. The analysis yielded a reliability coefficient of 0.92, which significantly exceeds the standard threshold of 0.70.

TABLE I. Summary of Dataset Characteristics (N=1,275)

Feature	Category	Frequency	Percentage (%)
Gender	Male	188	14.80%
	Female	1087	85.20%
School Type	Central Elementary School	847	66.22%
	Senior High School	123	9.69%
	Junior High School	305	24.02%
Professional Rank	Master Teacher	71	5.51%
	Teacher 1	909	71.34%
	Teacher 2	112	8.82%
	Teacher 3	170	13.31%
	Job Order	2	0.16%
	School Head/School School Principal	9	0.71%
	Volunteer Teacher	1	0.08%
	SpEd LGU	1	0.08%
Education	BS	841	65.83%
	MS	399	31.42%
	PhD	27	2.13%
	Post Doctorate	8	0.63%
Reliability	**Cronbach's Alpha**	**0.92**	**Excellent**

B. Data Pre-Processing

After data acquisition, the data is pre-processed. It was determined that there are 11 to 20 missing instances across several high-dimensional features. The Mode Imputation was applied, an imputation strategy that replaces missing values with the most frequent value, to address missing categorical and ordinal variables.

Handling missing data includes checking for missing values in the dataset and applying appropriate techniques such as imputation to fill in missing values. Data transformation was subsequently applied, which covers converting categorical variables to numerical format if necessary.

C. Theoretical Frameworks

- The theoretical foundation of this study is anchored in the European Framework for the Digital Competence of Educators (DigCompEdu) in Fig. 4 from [4].
- Digital competence [1] is defined not as a singular skill, but as a composite of six interrelated Latent Constructs, each representing a specific functional layer of the educational system:
- Professional Engagement: Focused on the organizational communication (PCO) and collaborative practices (PCC) of the educator.
- Digital Resources: The technical capacity to source, create (DRC), and manage digital assets.
- Digital Pedagogy: The core operational layer, encompassing the implementation of digital tools in teaching (DPT) and learner guidance (DPG).
- Assessment (Evaluation): The analytical layer where educators use digital technologies to enhance assessment strategies (EFA) and analyze evidence to provide feedback.
- Empowering Learners: The inclusive layer. It measures how technology is used to support learner diversity, accessibility (ESA), and personalized learning (ESD).

The final phase of the framework involved Predictive Analysis using a suite of supervised learning algorithms: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest. Accuracy calculations and predictive classification of digital divide levels were conducted for all models, with comparative performance metrics utilized to select the optimal classifier for the regional dataset.

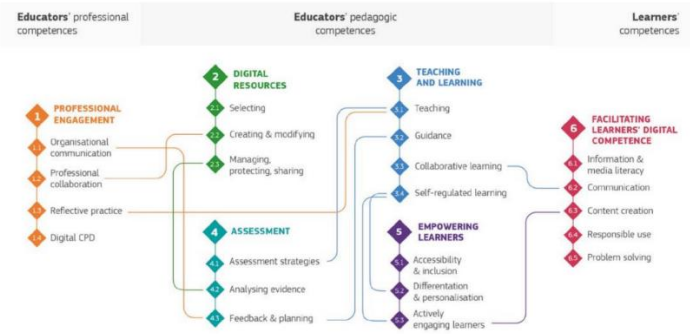


Fig. 4. Illustration of the digcompedu framework [4].

D. Structural Equation Modeling

The hybrid framework provides a two-stage approach to analyzing teacher digital competency in Fig. 5. Phase 1 uses Structural Equation Modeling (SEM) to validate theoretical constructs, identifying "Professional Engagement" as the system's foundational engine. These validated factors then bridge into Phase 2, which employs an SMOTE-ENN machine learning pipeline to handle imbalanced data and generate predictive proficiency classifications. The architecture demonstrates that an integrated approach—combining causal validation with high-performance algorithms like Random Forest (89% accuracy)—is more effective than isolated analytical methods for diagnosing and addressing pedagogical gaps in regional educational systems.

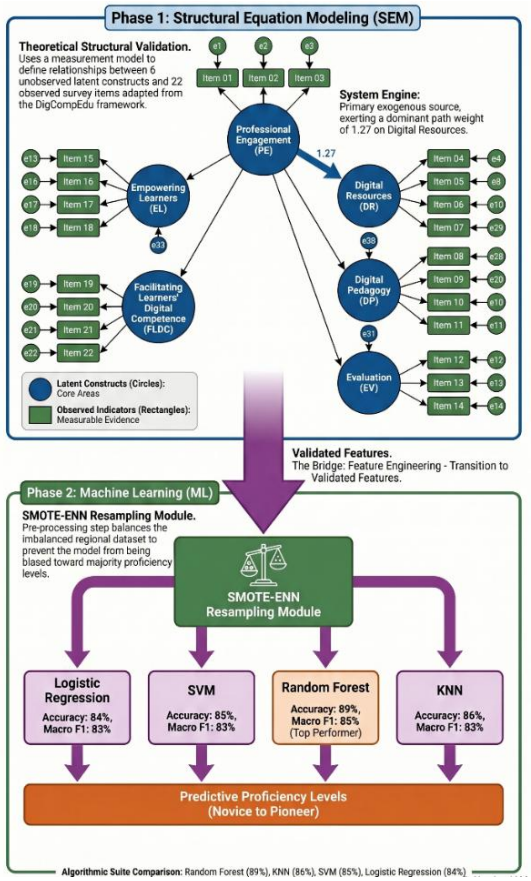


Fig. 5. Proposed model architecture.

The structural validation of the theoretical framework was conducted using Structural Equation Modeling (SEM) executed in SPSS AMOS. The model was specified as a full structural equation system comprising two distinct components: the measurement model and the structural model. The measurement model defined the relationship between six latent constructs (unobserved variables) and 22 observed indicators (survey items). Specifically, the constructs Professional Engagement, Digital Pedagogy, Digital Resources, Evaluation, Empowering Learners, and Facilitating Learners' Digital Competence were mapped to their respective observed variables. The Structural Model is the Directional paths that were established to represent hypothesized causal relationships. Professional Engagement was specified as the exogenous variable, while the remaining five constructs were specified as endogenous variables.

To account for measurement bias and unexplained variance, the model incorporated first Measurement Errors (e_1 through e_{22}), which are assigned to each observed indicator to capture

variance specific to the individual survey items. Second Structural Disturbance Terms ($e_{29}, e_{30}, e_{31}, e_{33}, e_{34}$) are assigned to each endogenous latent construct to represent the residual variance not accounted for by the structural paths in the system.

The model was constrained to ensure mathematical identification and a stable metric scale. For each latent construct, one regression weight to an observed indicator was fixed to 1.0 (e.g., the paths to PCD, DPS, DRM, EFF, ESAS, and FSDP). This procedure transformed the latent constructs into a standard metric, allowing the system to estimate the remaining free parameters and path coefficients.

$$F_{ML} = \ln|\Sigma(\theta)| + \mathrm{tr}(S\Sigma^{-1(\theta)}) - \ln|S| - p \quad (1)$$

The Maximum Likelihood Estimation (MLE) algorithm in Eq. (1) was utilized to calculate the path coefficients and variance estimates.

The structural integrity of the model was evaluated against a suite of computational fit indices [36]. Parsimony Assessment for CMIN/DF ratio (also known as the Normed Chi-Square) is a measure used to evaluate the "parsimony" or the efficiency of the model [36]. Absolute Fit Testing for Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), and Adjusted Goodness of Fit Index (AGFI) were employed, alongside Incremental Fit Comparison for Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Normed Fit Index (NFI) [36]. Information Criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), were also utilized for model comparison [36]. However, contemporary literature cautions against the use of rigid, fixed cutoff values for these indices, recommending instead that researchers consider model-specific characteristics and context [37][38].

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i}} \quad (2)$$

Lastly, in the final phase, the validated features were utilized to construct a predictive classifier using Logistic Regression in Eq. (2). To ensure the reliability of the predictive results and prevent model overfitting, the dataset (N=1,276) was partitioned using a train(80%)-test(20%) split. To validate the stability of the classifier across different subsets of the regional data, a 5-Fold Cross-Validation procedure was executed.

The predictive accuracy of the model was quantified using a comprehensive suite of classification metrics derived from the Confusion Matrix first is Accuracy - the overall ratio of correctly predicted instances to total instances. Precision and Recall are utilized to evaluate the model's ability to correctly identify positive instances of high ICT competency while minimizing false positives and false negatives. F1-Score - the harmonic mean of precision and recall, serving as a singular metric for balanced classification performance.

E. Class Imbalance

The original dataset exhibited a significant class imbalance, which could bias the predictive models toward majority classes. To mitigate this, a hybrid SMOTE-ENN (Synthetic Minority Oversampling Technique and Edited Nearest Neighbors) pipeline was applied to the training set (N=891). SMOTE was utilized to oversample minority classes by generating synthetic instances, while ENN was subsequently applied to remove ambiguous or 'noisy' samples from the feature space, ensuring well-defined decision boundaries. Table II illustrates the proficiency level distribution before and after the resampling process. The results confirm that this hybrid strategy successfully redistributed the instances to provide a more equitable analytical foundation for the machine learning models.

TABLE II. DISTRIBUTION OF PROFICIENCY LEVELS BEFORE AND AFTER SMOTE-ENN

Proficiency Level	Count (Before SMOTE-ENN)	Count (After SMOTE-ENN)
Pioneer	25	317
Novice	58	315
Leader	109	283
Explorer	328	201
Integrator	468	122
Expert	287	98
Total	**891**	**1,336**

IV. THE STRUCTURAL EQUATION MODEL

A. Goodness-of-Fit (GOF) and Model Verification

To determine the extent to which the theoretical architecture matches the empirical data collected from the 1,276 regional educators, the model was subjected to a rigorous Goodness-of-Fit (GOF) assessment. The observed covariance matrix was compared against the model-implied matrix using Maximum Likelihood Estimation in Eq. (1).

TABLE III. CALCULATED GOODNESS-OF-FIT FOR THE STRUCTURAL MODEL

Fit Index Category	Metric	Calculated Value	Interpretation
Parsimony Fit	CMIN/DF	4.749	Acceptable (Threshold < 5.0)
Absolute Fit	GFI	0.935	Good Fit
	AGFI	0.917	Good Fit
	RMSEA	0.054	Excellent (Threshold < 0.06)
Incremental Fit	NFI	0.936	Strong Fit
	CFI	0.948	Strong Fit
	TLI	0.940	Strong Fit

Table III shows that the model achieved convergence with high reliability. The RMSEA of 0.054 is a critical result, indicating that the structural model possesses a very low discrepancy per degree of freedom. This validates the model's generalizability across the diverse demographics of the Bukidnon teacher population. Furthermore, the GFI (0.935) and CFI (0.948) scores confirm that the proposed six-construct structure accounts for the majority of the variance in the observed indicators.

The modeling procedure followed a multi-stage mathematical specification to assess the inter-dependencies between latent constructs in Fig. 3.

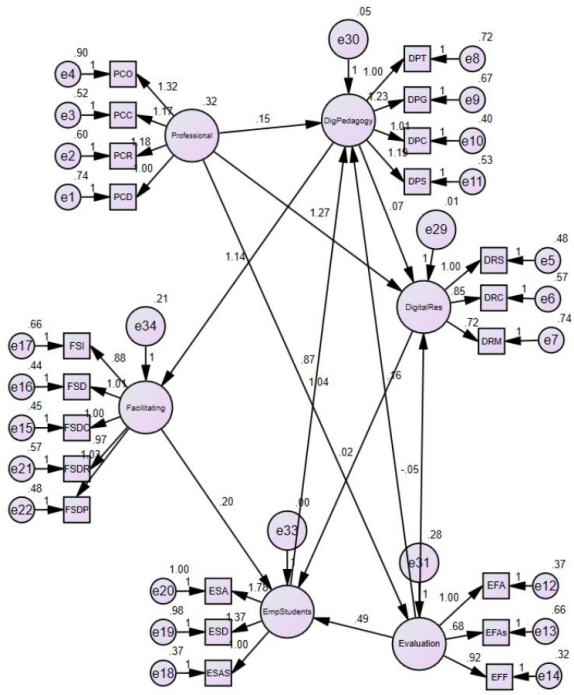


Fig. 6. SEM framework implemented in SPSS AMOS.

B. Path Analysis and Predictive Weights

Following the global validation of the model, the individual structural paths were analyzed to identify the primary drivers of teacher competency. Fig. 6 illustrates the final structural diagram with the calculated unstandardized regression weights.

The path analysis reveals a clear hierarchical influence within the digital competency system:

1) System driver: The path from Professional Engagement to Digital Resources yielded the highest weight in the system (1.27). This indicates that for every one-unit increase in professional commitment, the competency in sourcing and creating digital resources increases by 1.27 units, identifying "Professional Engagement" as the foundational engine of the system.

2) Pedagogical integration: A significant structural relationship was confirmed between *Evaluation and Empowering Students* (0.49). This confirms that data-driven assessment practices are essential precursors to achieving the ultimate inclusive goal of student digital empowerment.

3) Direct influence: The model shows that *Professional Commitment* also maintains a direct positive influence on Digital Pedagogy (0.15), suggesting that commitment to reflective practice directly modulates how tools are applied in the classroom.

C. Structural Validity Conclusion

The results of the SEM phase demonstrate that the digital divide is not a set of isolated skills but a highly integrated system. The high GOF scores prove that the DigCompEdu framework is an appropriate mathematical model for the Philippine regional context. Meeting all statistical thresholds, the latent constructs—specifically Digital Pedagogy and Professional Engagement—are validated as high-quality features. This structural certainty provides the necessary evidence base for the final Machine Learning Predictive Analysis presented in the following section.

V. Machine Learning Predictive Results

Following the structural validation of the constructs, the predictive phase was executed using the resampled dataset. The application of the SMOTE-ENN (Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors) pipeline was utilized to ensure that the classifiers were not biased toward the majority proficiency levels, thereby providing an equitable analysis of the digital divide.

A. Comparative Model Performance

The predictive utility of the four supervised learning algorithms was evaluated based on Accuracy, Macro-averaged Precision, Recall, and the F1-Score. Table IV summarizes the performance of the models within the optimized SMOTE-ENN framework.

TABLE IV. Predictive Performance Metrics with SMOTE-ENN Resampling (N=1,276)

Model	Accuracy	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1-Score
Logistic Regression	84%	84%	81%	83%
SVM	85%	84%	83%	83%
Random Forest	**89%**	86%	84%	85%
K-Nearest Neighbors	86%	84%	82%	83%

B. Algorithmic Analysis and Model Selection

As illustrated in Table IV, all four classifiers demonstrated high predictive performance, with F1-scores consistently exceeding 83%. The Random Forest (RF) algorithm emerged as the best performer, achieving an overall accuracy of 89% and a Macro F1-Score of 85%. The superior performance of RF is attributed to its ensemble nature, which effectively aggregates multiple decision trees to capture the non-linear relationships between the pedagogical and professional features identified in the SEM phase.

The K-Nearest Neighbors (KNN) and SVM models followed closely with accuracies of 86% and 85%, respectively. The high recall scores across these models (up to 84%) indicate that the SMOTE-ENN strategy successfully enabled the system to identify minority proficiency classes (such as "Novice" and "Leader"), which are often missed in traditional imbalanced models.

The Logistic Regression model, while having the lowest accuracy at 84%, still provided significant interpretive value through its 81% recall rate. The narrow margin of difference between the four models suggests a high degree of consensus regarding the predictive importance of the structural features used as inputs.

C. Confusion Matrix Evaluation

To visually verify the classification accuracy, a composite confusion matrix analysis was performed for each algorithm in Fig. 7.

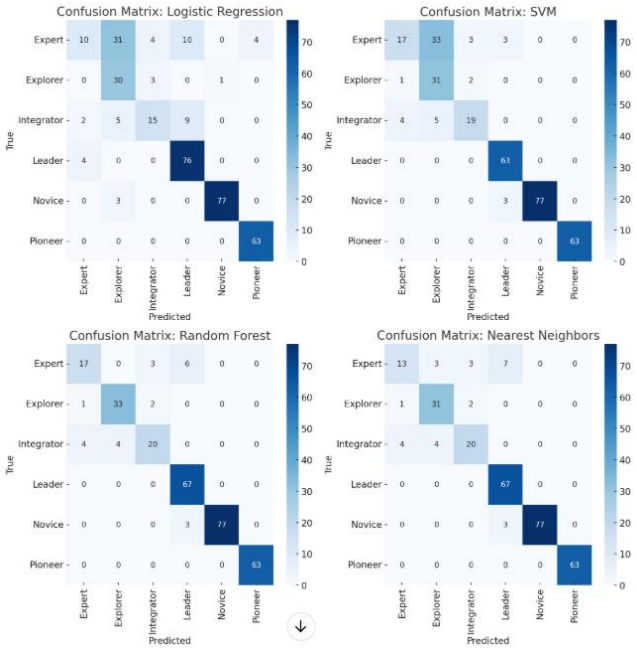


Fig. 7. Grid of Confusion Matrices for optimized models.

The diagonal values in Fig. 7 confirm a high rate of True Positives across all proficiency levels. The low off-diagonal counts demonstrate that the integrated SEM-ML framework is highly effective at minimizing misclassification errors, providing a reliable diagnostic tool for assessing the digital divide in regional education systems.

D. Feature Importance and Predictive Determinants

To enhance the transparency of the predictive framework and identify the primary drivers of the digital divide, a feature importance analysis was conducted. This stage translates the "black-box" predictions of the algorithms into actionable insights for educational policy.

1) Identifying the strongest predictors: Based on the coefficients derived from the Logistic Regression model and the feature importance rankings of the Random Forest classifier, Digital Pedagogy in Teaching (DPT) emerged as the most significant predictor of teacher ICT competency. The model yielded an Odds Ratio (OR) of 1.5176 for this feature, indicating that for every unit increase in a teacher's pedagogical integration score, the likelihood of belonging to a higher proficiency category increases by 51.76%.

2) The role of digital resource creation: The analysis further identified Digital Resource Creation and Modification (DRC) as a critical secondary predictor (OR = 1.3945). This aligns with the SEM results (Path weight = 1.27), confirming that the ability to customize digital content is a structural requirement for moving beyond the "Novice" stage of the digital divide.

3) Inverse predictors and system gaps: A compelling and critically revealing finding is the inverse relationship between Highest Educational Attainment (HighED) and ICT competency (OR = 0.2879). This negative coefficient indicates that, within this regional population, traditional advanced academic credentials do not serve as a proxy for digital proficiency. This "Academic-Digital Mismatch" highlights a systemic gap: postgraduate and teacher-education curricula in the region may be prioritizing traditional theoretical frameworks over the applied digital-pedagogical competencies defined by the DigCompEdu framework. This suggests that senior, highly-credentialed educators are not necessarily more digitally prepared; rather, they may be the most at risk of falling behind in the digital transition if their academic training is not synchronized with modern instructional technology requirements.

4) Strategic implications: The feature importance results demonstrate that the digital divide in Bukidnon is primarily pedagogical rather than technical. While "Professional Collaboration" and "Organizational Communication" are foundational (as proven in the SEM phase), the Machine Learning models prove that actual classroom implementation (Pedagogy) is the variable that ultimately decides a teacher's proficiency level.

VI. Discussion

A. Characterizing the Digital Divide in Bukidnon

The integration of SEM path analysis and Machine Learning feature importance reveals that the digital divide among teachers in Bukidnon is not a binary gap of "haves vs. have-nots," but a multi-layered structural bottleneck.

1) The SEM result identifying professional engagement as the strongest driver (Path weight = 1.27) suggests that regional teachers possess the necessary organizational commitment and reflective drive to bridge the divide. However, the high variance in the Digital Resources construct indicates that this commitment does not always translate into resource access.

2) While SEM identifies commitment as the driver, the Machine Learning results identify Digital Pedagogy in Teaching (DPT) as the ultimate predictive determinant (OR = 1.5176). This synthesis suggests that while teachers are committed, the "Digital Divide" is actually a Pedagogical Divide. An educator's ability to cross into the "Expert" or "Leader" categories depends almost entirely on their capacity to integrate tools into instructional design rather than simply possessing the tools themselves.

3) The surprising inverse relationship between Highest Educational Attainment and competency (OR = 0.2879) highlights a significant regional gap. It suggests that traditional advanced academic training in the province has not yet integrated the digital competencies defined by the DigCompEdu framework, creating a scenario where senior, highly credentialed teachers may be the most at risk of falling behind in the digital transition.

B. Theoretical Implications for the DigCompEdu Framework

The study provides a successful large-scale validation (N=1,276) of the DigCompEdu framework in a predominantly agricultural, regional context.

The high global fit indices (CFI=0.948, GFI=0.935) confirm that the framework's six latent domains are cross-culturally applicable and mathematically stable even in resource-constrained environments. The structural analysis confirms the framework's theoretical assumption that *Professional Engagement* must precede *Digital Pedagogy*.

However, our findings suggest that in regional settings, the path between *Professional Commitment* and *Digital Resources* is more volatile than in urban settings, requiring the framework to perhaps place more weight on "Institutional Context" when applied to regional clusters.

C. Technical Contribution

A primary technical contribution of this research is the demonstration of the Hybrid SEM-ML Analytical Pipeline. In typical Machine Learning studies, features are often fed into models without conceptual validation. This study uses SEM as a pre-filtering layer, ensuring that only constructs with high structural validity and low measurement error are used for prediction. This reduces "noise" and increases the reliability of the classification.

From an Information Systems perspective, the success of the Random Forest model with 89% accuracy proves that SMOTE-ENN is an essential pre-processing layer for regional educational data. Balancing the sparse "Novice" and "Leader" classes, the system achieved a Macro F1-Score of 85% that far exceeds traditional linear methods.

This hybrid approach provides a blueprint for other regional government agencies to move from descriptive "snapshot" surveys to Predictive Diagnostic Systems. The model can be used to develop Decision Support Systems (DSS) that automatically identify teachers who require specific pedagogical interventions based on their structural profile.

VII. Conclusion and Recommendations

A. Final Conclusion

This research successfully developed and validated a Hybrid Analytical Framework integrating Structural Equation Modeling (SEM) and Machine Learning (ML) to investigate the digital divide among 1,276 public school teachers in Bukidnon, Philippines. The study concludes that the digital divide is not merely a consequence of resource scarcity but a complex system of structural interdependencies.

The SEM phase established the mathematical validity of the DigCompEdu framework in a regional context, proving that Professional Engagement acts as the primary engine driving digital integration. The subsequent Machine Learning phase achieved a high predictive accuracy of 89% using the Random Forest algorithm, identifying Digital Pedagogy in Teaching (DPT) as the most critical predictive determinant of an educator's proficiency level.

Furthermore, the surprising inverse relationship between formal academic attainment and ICT competency highlights a critical mismatch between traditional higher education and the practical digital skills required in the modern classroom. Ultimately, this study demonstrates that a data-driven, hybrid approach provides the precision necessary to move beyond descriptive reports into predictive diagnostics for educational systems.

B. Study Limitations and Future Directions

While this research provides a robust hybrid SEM-ML framework for diagnosing digital competency, several limitations should be considered. First, the data were collected using a self-report Likert-scale instrument, which may be susceptible to social desirability and self-report bias; participants might overestimate their digital proficiencies. Second, the study's geographic scope is restricted to a single division (Bukidnon), which may limit the generalizability of the findings to broader or more urbanized regional contexts. Finally, the study employs a cross-sectional research design. While Structural Equation Modeling (SEM) identifies significant structural pathways, this design cannot definitively establish the causal directionality of these relationships. Future research should prioritize longitudinal studies to measure the evolution of teacher digital competency over time and incorporate objective performance data alongside self-reported metrics to further validate these findings.

C. Recommendations for Educational Policy

Based on the validated predictive weights of the model, the following interventions are recommended for the Department of Education (DepEd) and regional stakeholders:

1) Prioritize pedagogical integration over technical skills: Professional development programs should pivot from basic tool-use training to Digital Pedagogy. Training should focus on how to integrate ICT into instructional design, as this factor has the highest odds ratio for moving teachers out of the "Novice" and "Explorer" categories.

2) Bridge the 'academic-digital mismatch' for highly credentialed educators: The negative correlation between formal advanced degrees and digital competency points to a significant alignment gap between traditional Higher Education Institutions (HEIs) and the practical realities of the digital classroom. To resolve this, the Department of Education (DepEd) should initiate a collaborative curriculum audit with regional HEIs to integrate applied DigCompEdu modules into postgraduate programs. Furthermore, we recommend specialized, track-based upskilling programs for senior teachers and those with advanced degrees. These interventions should focus on moving these educators from theoretical academic mastery toward "Expert" and "Leader" levels in digital pedagogical implementation, ensuring that their tenure and qualifications are effectively leveraged to drive digital transformation.

3) Institutionalize collaborative learning communities: Given the foundational role of Professional Engagement in the SEM results, schools should establish formal digital mentorship

programs that reward collaborative resource creation and reflective practice.

D. Technical Recommendations and Future Work

From a computational perspective, this study provides a blueprint for regional Data for Development initiatives.

Future research directions include system automation, algorithm expansion, and longitudinal modeling. System Automation involves integrating the validated SEM-ML pipeline into a web-based Decision Support System (DSS). This would allow educational administrators to input survey data and automatically receive real-time "Proficiency Risk" scores for their teaching staff.

For Algorithm Expansion, while Random Forest and Logistic Regression proved effective, future studies should explore Deep Learning architectures, such as Artificial Neural Networks (ANN), to see if predictive accuracy can be further enhanced for larger national-level datasets.

Longitudinal Modeling: Future work should utilize the proposed framework in a longitudinal study to measure the rate of competency growth following specific training interventions, allowing for the calculation of the "Return on Investment" (ROI) for digital transformation policies.

ACKNOWLEDGMENT

We sincerely thank the Department of Education, Bukidnon, for helping us to conduct this research. We would also like to thank Central Mindanao University for funding our research Project with a code, R-0355, which allowed us to conduct our study about ICT proficiency.

DECLARATION OF GENERATIVE AI IN SCHOLARLY WRITING

During the preparation of this manuscript, the authors utilized Google AI Studio to enhance the clarity, flow, and overall readability of the text. Following the use of this tool, the authors reviewed, validated, and edited the content to ensure accuracy. The authors take full responsibility for the integrity of the information presented. No AI tools were employed for the purposes of data analysis, hypothesis testing, or the derivation of research insights.

ETHICAL DECLARATIONS

This study was conducted in accordance with the ethical standards of Central Mindanao University and was formally approved by the Department of Education (DepEd) Division of Bukidnon. Research ethics clearance was secured from the Institutional Ethics Review Committee before the commencement of data collection. Informed consent was obtained from all 1,276 participants. To ensure participant protection, all data were anonymized and processed to maintain strict respondent confidentiality throughout the research process.

REFERENCES

[1] E. Onsay and J. Rabajante, "Measuring the Unmeasurable through Machine Learning Regressions and Classifications: Multidimensional Poverty Predictions in the Poorest Region of Luzon, Philippines," 2024.

[2] C. S. R. Saflor, "Modeling Student Acceptance of AI Technologies in Higher Education: A Hybrid SEM–ANN Approach," Future Internet, vol. 17, no. 12, p. 581, 2025.

[3] M. S. Asagar, "Digital competence in education: A comparative analysis of frameworks and conceptual foundations," Synergy: Int. J. Multidiscip. Stud., vol. 2, no. 1, pp. 9–23, 2025. Online]. Available: https://sijmds.com/index.php/pub/article/view/34

[4] J. Mattar, C. C. Santos, and L. M. Cuque, "Analysis and comparison of international digital competence frameworks for education," Education Sciences, vol. 12, no. 12, p. 932, 2022.

[5] V. Dolgopolovas and V. Dagiene, "Competency‐based TPACK approaches to computational thinking and integrated STEM: A conceptual exploration," Comput. Appl. Eng. Educ., vol. 32, no. 6, p. e22788, 2024.

[6] L. Sabyrkhanova et al., "Digital Literacy as a Tool for Developing Computational Thinking in Young Learners: A Literature Review," Open Educ. Stud., vol. 8, no. 1, p. 20250121, 2026.

[7] C. Schreiner and C. Wiesner, "Modelling digital competence by combining computational thinking with general learning taxonomies," Eur. Educ. Res., vol. 6, no. 1, pp. 21–42, 2023.

[8] S. K. Softić, "Teachers' digital competences as a key factor for the digital transformation of education," Adv. Online Educ., vol. 1, no. 1, pp. 75–86, 2022.

[9] J. Abella and E. D. Rosa, "Digital literacy and digital competence of selected Filipino teachers: Basis for a post-pandemic pedagogy," 2023.

[10] M. R. Briones et al., "e-Work Self-Efficacy digital competence and work engagement of teachers in public secondary schools from two provinces in the Philippines," Ho Chi Minh City Open Univ. J. Sci.-Soc. Sci., vol. 14, pp. 14–36, 2023.

[11] C. A. Velandia Rodríguez, A. F. Mena-Guacas, S. Tobón, and E. López-Meneses, "Digital teacher competence frameworks evolution and their use in Ibero-America up to the year the COVID-19 pandemic began: a systematic review," Int. J. Environ. Res. Public Health, vol. 19, no. 24, p. 16828, 2022.

[12] R. S. Alotaibi and S. M. Alshahrani, "An extended DeLone and McLean's model to determine the success factors of e-learning platform," PeerJ Comput. Sci., vol. 8, p. e876, 2022.

[13] A. F. Ali, R. H. Abdullah, A. Abi Hassan, H. O. Abdullahi, and M. M. Mohamed, "COVID-19 pandemic impact on e-learning adoption and its utilization at higher education: A comparative analysis of institutions and students' perspectives," Ingénierie des Systèmes d'Information, vol. 29, no. 2, p. 447, 2024.

[14] J. C. Vázquez-Parra, L. C. Henao-Rodriguez, J. P. Lis-Gutiérrez, I. M. Castillo-Martínez, and P. Suarez-Brito, "eComplexity: validation of a complex thinking instrument from a structural equation model," Frontiers in Education, vol. 9, p. 1334834, Jun. 2024.

[15] X. Ju, "Unveiling the Structure of Teacher Competency: An Eight-Dimensional Model for Primary Education," in Proc. 2025 21st Int. Conf. Comput. Intell. Security (CIS), Dec. 2025, pp. 1–5.

[16] M. K. Aydin, T. Yildirim, and M. Kus, "Teachers' digital competences: a scale construction and validation study," Frontiers in Psychology, vol. 15, p. 1356573, 2024.

[17] F. Liaquat and M. S. Farooq, "Development and Validation of a Digital Competence Scale for Teachers at Higher Education level," Acad. Educ. Soc. Sci. Rev., vol. 5, no. 3, pp. 348–358, 2025.

[18] M. Nowak and R. Zajkowski, "An Integrated Structural Equation Modelling and Machine Learning Framework for Measurement Scale Evaluation—Application to Voluntary Turnover Intentions," AppliedMath, vol. 5, no. 3, p. 105, 2025.

[19] J. S. Moreira-Choez et al., "Machine Learning Algorithms to Predict Digital Competencies in University Faculty," F1000Research, vol. 14, p. 573, 2025.

[20] L. Yang, Q. Wang, B. Zheng, X. Li, X. Ma, and T. Wang, "Assessing Digital Teaching Competence: An Approach for International Chinese Teachers Based on Deep Learning Algorithms," Scalable Comput.: Pract. Exper., vol. 25, no. 1, pp. 495–509, 2024.

[21] R. Benzer, S. Benzer, and T. Doğan, "Assessing science teachers' digital literacy with survey-based machine learning," Inf. Dev., 2025, doi: 10.1177/02666669251350669.

[22] D. Yan, X. Yuan, and G. Li, "Key Factors and Predictive Models of Digital Collaborative Education Based on Machine Learning," Annals of the New York Academy of Sciences, vol. 1555, no. 1, p. e70117, 2026.

[23] J. Cabero-Almenara, F. D. Guillen-Gamez, J. Ruiz-Palmero, and A. Palacios-Rodríguez, "Classification models in the digital competence of higher education teachers based on the DigCompEdu Framework: logistic regression and segment tree," J. E-Learn. Knowl. Soc., no. 1, pp. 49–61, 2021.

[24] C. Magasi, "Evaluating Machine Learning Approaches in Structural Equation Modelling to Improve Predictive Accuracy in Marketing Research," 2025.

[25] X. Li, Q. Qin, T. Zheng, and X. Wang, "Social Integration Among Rural-to-Urban Migrants: A Dynamic Analysis Using a Hybrid Approach Integrating Structural Equation Modeling and Machine Learning," Evaluation Review, vol. 50, no. 3, pp. 423–461, 2026.

[26] . M. Brandmaier and R. C. Jacobucci, "Machine learning approaches to structural equation modeling," in Handbook of Structural Equation Modeling, 2nd ed., R. H. Hoyle, Ed. New York, NY, USA: Guilford Press, 2023, pp. 722–739.

[27] L. Zhang, J. Gao, L. Zhao, Z. Liu, and A. Guan, "Predicting college student engagement in physical education classes using machine learning and structural equation modeling," Appl. Sci., vol. 15, no. 7, p. 3884, Apr. 2025, doi: 10.3390/app15073884.

[28] J. Doe, "Challenges and Opportunities in Higher Education Digital Transformation: A Combined SEM and Machine Learning Perspective," Research Square, Preprint, 2026. [Online]. Available: https://assets-eu.researchsquare.com/files/rs-7708749/v1_covered_fca8c5fb-cb17-4b35-a92f-1dcadd928f1d.pdf

[29] T. Wongvorachan, O. Bulut, J. X. Liu, and E. Mazzullo, "A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning," Information, vol. 15, no. 6, p. 326, 2024.

[30] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," Information, vol. 14, no. 1, p. 54, 2023.

[31] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," Comput. Educ.: Artif. Intell., vol. 6, p. 100222, 2024.

[32] G. Eom and H. Byeon, "Searching for optimal oversampling to process imbalanced data: Generative adversarial networks and synthetic minority over-sampling technique," Mathematics, vol. 11, no. 16, p. 3605, 2023.

[33] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," J. Big Data, vol. 12, no. 1, p. 71, 2025.

[34] R. Diallo, C. Edalo, and O. O. Awe, "Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score," in Pract. Stat. Learn. Data Sci. Methods: Case Stud. LISA 2020 Global Netw., Cham, Switzerland: Springer, 2024, pp. 283–312.

[35] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," Information, vol. 14, no. 1, p. 54, 2023.

[36] S. Sathyanarayana and T. Mohanasundaram, "Fit indices in structural equation modeling and confirmatory factor analysis: Reporting guidelines," Asian J. Econ. Bus. Account., vol. 24, no. 7, pp. 561–577, 2024.

[37] K. Groskurth, M. Bluemke, and C. M. Lechner, "Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions," Behav. Res. Methods, vol. 56, no. 4, pp. 3891–3914, 2024.

[38] D. McNeish and M. G. Wolf, "Dynamic fit index cutoffs for confirmatory factor analysis models," Psychol. Methods, vol. 28, no. 1, p. 61, 2023.

DeepEdgeNet: An Edge-Cloud Deep Learning Framework for Efficient Environmental Monitoring in IoT Systems

Qamar H. Naith
Department of Software Engineering-College of Computer Science and Engineering,
University of Jeddah, Jeddah 21959, Saudi Arabia

***Abstract*—Deep learning (DL) is currently considered one of the most powerful tools for environmental monitoring. Many environmental variables, such as air quality, climate, water, and energy, are monitored using Internet of Things (IoT) technologies. However, the DL-based environmental monitoring systems heavily depend on the cloud, and hence they suffer from latency, high energy consumption, and data privacy. This study proposes DeepEdgeNet, a distributed DL and Machine Learning (ML) framework for environmental monitoring systems based on edge computing and federated learning. In DeepEdgeNet, the features are extracted at the edge devices, and the model updates are aggregated at the central server without sharing the sensitive data. The proposed framework was evaluated on six IoT datasets collected from various environmental monitoring systems, such as air quality monitoring, household energy consumption, satellite-based land-cover classification, climate and extreme weather analysis, water quality assessment, and drought prediction. Experimental results have shown that the proposed system significantly improves the accuracy of the considered datasets, which are 94.5% for the Air Quality dataset, 95.4% for the EuroSAT dataset, and the mean absolute error (MAE) of time-series datasets is reduced up to 0.28 for drought prediction. Moreover, the proposed system has lower inference latency, up to 130 ms, and energy consumption compared with six state-of-the-art models. Although the edge–cloud environment was simulated using a unified experimental platform, the obtained results demonstrate the effectiveness of DeepEdgeNet for scalable and privacy-preserving IoT-based environmental monitoring applications. Hence, the proposed system is efficient and applicable to IoT-based environmental monitoring systems.**

Keywords*—Internet of Things (IoT); edge computing; federated learning; machine learning

I. Introduction

The world and the population change through the Internet of Things (IoT), which enhances the existing interaction with nature and its administration. These include sensors, weather stations, and monitoring systems that reveal real-time data related to primary environmental issues. The use of such technologies has increased concern for air pollution, flood forecasting, water resources, and biodiversity conservation. For instance, poisonous elements could be detected by air quality sensors, and water level sensors help predict floods, thus allowing authorities to act in disaster management quickly. However, increasing IoT deployments also results in ginormous data, so as to pose several challenges in processing data, energy consumption, and user privacy [1]. The pivotal role of such IoT-based systems in the facilitation of a few global sustainability goals is illustrated in Fig. 1.



Fig. 1. Key sustainability goals enabled by IoT systems.

Often, traditional ones rely on cloud processing, mostly as centralized processing for data harvested from IoT devices. However, despite having fit applications, the approach eventually stumbles in delay and bounces back and forth between the central servers. It seems like another good reason why the energy consumption of centralised systems can be very variable, and also the privacy issues related to sensitive data such as industrial emissions or urban air quality measurements. Centralised systems (e.g., using cloud servers) generally have latency and therefore cannot be applicable to real-time environmental monitoring systems such as flood monitoring systems or also real time air pollution alert systems for busy cities. Another key point is that in such environmental monitoring applications, the centralised systems have high bandwidth requirements. An illustration of the difference in location where the data processing takes place between a cloud computing system and an edge computing system is shown in Fig. 2. In cloud computing, all processing in the data server at the central location (cloud) has to take place in advance to ensure that the results come at a reasonable time, leading to a latency period between data reception and when any actions are taken in response to the environmental phenomena of interest. In addition, a large amount of bandwidth is required for the transfer of large amounts of data from numerous sensor nodes to a single central location (cloud). On the other hand, in the case of edge computing, the sensor node can start processing data immediately, thus a lower latency period occurs, and an overall

decrease in bandwidth consumption occurs due to data being processed locally. So, the transition to sustainable applications of IoT is another important transition from cloud computing to edge computing systems.

This study presents DeepEdgeNet, a novel framework dedicated to combining deep learning with edge computing to improve the efficiency and effectiveness of environmental IoT sensing applications. DeepEdgeNet processing stays along the edge of IoT networks, systematically performing real-time data analysis, thus making it less dependent on continuous communication with centralized servers. These combined qualities of latency and energy efficiency thus make it a suitable system for resource-limited environments [3]. This framework specifically comprises advanced deep-learning technologies, for example, Convolutional Neural Networks (CNNs), which work over spatial patterns -specific areas with high pollutant concentrations- and Temporal Fusion Transformers (TFTs), which utilize effective time-series data for temperature changes highlighted over weeks or even months. Finally, DeepEdgeNet integrates federated learning in which models can train across devices without raw data having an opportunity to get to a common point in ensuring privacy and security.

The framework talks about other doctrinal environmental threats such as projected air quality and its trends; the risk of floods based on rainfall and water level data; and tracking changes in biodiversity in protected areas. Notably, DeepEdgeNet was tested against six popular IoT datasets concerned with air quality, climate, and weather, along with biodiversity and other urban manifestation indices. It has been compared with performance across six other leading algorithms, including Spatio-Temporal Dynamic Graph Neural Networks (ST-DGNN) and Multi-Scale Convolutional Neural Networks (MSCNN). The evaluation considered prediction accuracy, processing speed, energy efficiency, and error rates as parameters for comparison.

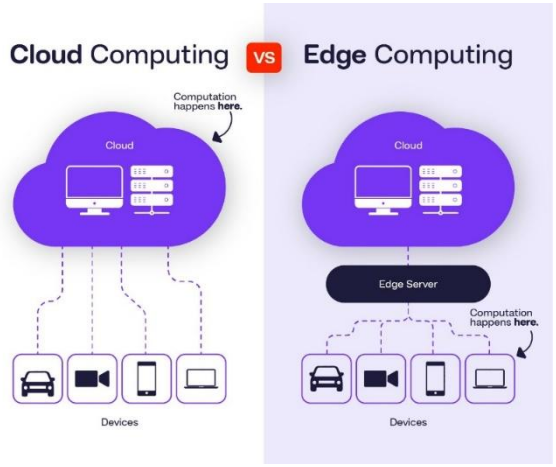


Fig. 2. Comparison between traditional cloud computing and edge computing.

This research work contributes in many ways: First, it presents a highly scalable and energy-efficient framework that IoT and edge computing can exploit to mitigate various environmental issues. Second, this study shows the flexibility and efficacy of the system by testing it with various real datasets under real-world scenarios. Third, it is shown that DeepEdgeNet outperforms all other methods with respect to accuracy and speed while consuming less energy. DeepEdgeNet also provides a pragmatic solution to environmental monitoring by leveraging existing IoT infrastructures that are easily adaptable to rapidly respond to challenges such as air pollution, floods, and biodiversity loss. Overall, the DeepEdgeNet makes important progress in environmental sensor networks based on IoT. Being an advanced technology, it can be applied in the field of deep learning and further research into processing power among decentralized edge computing to provide a versatile and efficient approach to complex environmental systems management.

The rest of this study is organized as follows: Section II provides a comprehensive review of related work, highlighting recent advancements in IoT-based environmental monitoring and the state-of-the-art algorithms used for comparison. Section III introduces the proposed DeepEdgeNet framework, detailing its system architecture, technical features, and implementation. Section IV describes the six IoT datasets utilized in the study and the evaluation metrics used to assess performance. Section V presents the experimental setup. Section VI shows the experimental results, including a detailed comparison of DeepEdgeNet with six state-of-the-art algorithms across the datasets. Finally, Section VII concludes the study, summarizes the key findings, and outlines future directions to further enhance the capabilities of the framework.

II. Related Work

Several changes have occurred in terms of environment monitoring and sustainability due to the advancement of IoT and edge computing.

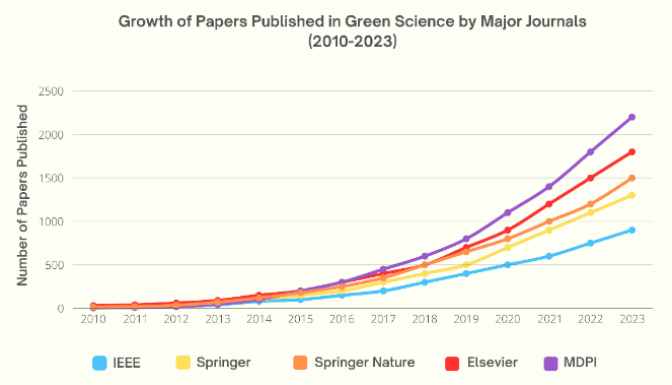


Fig. 3. Growth of papers published in green science by major journals (2010-2023).

Fig. 3 shows how research has grown in the field of green science and sustainability, as proclaimed by the leading journals like IEEE, Springer, Elsevier, MDPI, and Springer Nature. This section synthesizes insights from important state-of-the-art works in the context of their contribution towards IoT, edge computing, and machine learning techniques for environmental monitoring and sustainability. Chen S. et. al. present IoT-based smart grids with edge computing for energy consumption optimization in real-time. Distributed systems are significantly shown to be speedier and more efficient in the workings of the grid and lead to fewer carbon emissions [3]. Mobile edge computing has also been investigated in another study related to

IoT applications for showing the benefits of reduced bandwidth and latency usage, corresponding to the architecture of DeepEdgeNet for on-the-fly environmental monitoring purposes [2].

Omniawa et. al article was comprehensive and detailed in surveying all the possible fog and edge computing frameworks applied to IoT-based applications. It also discussed issues affecting such frameworks, like scalability and privacy, and proposed some hybrid architectures to mitigate those concerns [4]. The same study depicts some evidence in favor of edge computing and its usefulness in IoT-based manufacturing, reduced time delay, and efficient processing that are in line with DeepEdgeNet targets for lowering latencies in environmental monitoring [5].

Okot et. al. explore the potential of IoT in achieving sustainability goals in Costa Rica by providing a conceptual framework. The study highlights how IoT plays an important role in achieving energy efficiency, environmental monitoring, and resource management by addressing several scalability and integration challenges. Costa Rica's promise of utilizing renewable energy and conservation efforts gives it a strong case for rolling out IoT ecosystems for sustainable practice promotion [6]. Researching intelligent edge computing for energy management in smart cities demonstrates how AI algorithms can be adopted in predicting energy consumption and optimized usage; it lists applicable combinations alongside DeepEdgeNet capabilities, such as flooding prediction or air quality monitoring [7]. Furthermore, an SDN-based edge for IoT-enabled health care systems study underscores the critical aspects of managing secure data and resource optimization, mostly relevant to managing sensitive environmental data [8].

According to Ramadane et al., a new approach for assessing the environmental effect of IoT objects has been developed using graph-based machine learning models. The research represents IoT devices as graphs, thus simplifying traditional resource-intensive Life Cycle Assessments (LCAs) to ensure ecological sustainability with fast-growing IoT [9]. Regarding the environment, different IoT frameworks have been used to monitor air quality, biodiversity, and the sustainability of urban areas. For example, real-time air quality assessment systems with IoT use deep learning to precisely identify pollution trends in terms of real-time monitoring of pollutants and their forecasting [10]. Other works focus focused on IoT applications towards biodiversity conservation as well as the need for understanding spatio-temporal data to monitor the distribution of species [11]. About Narayana et al. [12] smart environmental monitoring system built with a range of IoT sensors (for temperature, humidity, atmospheric pressure, and hazardous gas activities) that keep checking conditions and trigger alerts in case a threshold is exceeded. So this type of real-time IoT system could mitigate risks and prevent disasters by reporting environmental hazards to users.

Piramuthu S. investigates how IoT technologies can support environmental sustainability in agricultural supply chains. It highlights IoT's role in reducing food waste, optimizing resource consumption, and enabling precision farming. From "farm to fork," IoT aids in monitoring crops, managing resources, and improving supply chain efficiency to reduce environmental footprints [13]. Nirere et al. demonstrate the establishment of a cost-effective, real-time IoT-based climate prediction system deployed in the industrial economic zone of Rwanda itself. The system uses integration of sensors such as MQ-135 and DHT11 to monitor environmental parameters such as CO2 levels, temperature, and humidity. The collected data is processed using MQTT and stored on MongoDB as well as used in machine learning models for analysis and prediction. The study evaluates different algorithms related to machine learning, which include Logistic Regression, Random Forest, Gradient Boost, Decision Tree, K-Nearest Neighbor, and Multilayer Perceptron. In accuracy, Random Forest surpassed the case with 99% training accuracy and 84% test accuracy. Results from the study showed that the high levels of CO2 and increase in temperature have manifested that, above these interventions, the warming periods are going to dominate in the future. This research opens several opportunities for IoT and machine learning in making actionable insights toward climate change mitigation and adaptation [14].

Federated Learning is supposed to be one of the techniques that shows great promise in terms of privacy preservation and scalability in applications within distributed IoT systems. For instance, the work by [15] presented a framework that was an edge-based FL adaptable framework on IoT contexts that could well be used to allow millions of devices to come together and be able to train their models without processing their raw data in the cloud. The proposed program, using the edge nodes for training and simply modular upgrades, reduced congestion on the network and protected the user's privacy – basically, the DeepEdgeNet building block. In addition, there are further applications of FL in the Industrial IoT in terms of efficiency increase, such as in the work by Hsu et al. [16], which utilizes a federated learning scheme to sample data from IIoT networks. The procedure significantly reduces redundant transmissions and, alongside a very small model accuracy reduction, provides a considerable amount of energy savings. Hence, it shows how connected learning can increase battery life in sensors and also scale to large deployments. These types of strategies, based on FL, prove that a theoretical understanding of distributed learning in IoT can practically benefit systems due to data local retention, a principle that DeepEdgeNet leverages for privacy and scalability in environmental monitoring.

Qamar et al. emphasize the transformative potential of IoT in environmental research and resource management. It outlines IoT's applications in real-time environmental monitoring, smart agriculture, and water resource management, focusing on efficiency, innovation, and sustainability. The study also identifies challenges such as data privacy and implementation costs [17]. Abd El-Mawia et al. critique the current limitations of climate change mitigation strategies and propose IoT as a transformative tool for adaptation. It focuses on Industrial IoT (IIoT) as a future solution, emphasizing its role in monitoring and reducing greenhouse gas emissions, enabling smart energy systems, and fostering sustainable industrial practices [18].

The following six algorithms are used for the comparative evaluation of DeepEdgeNet:

- Temporal Fusion Transformer (TFT), as it is built, is focused on applying different time-series analyses to

different types of time series. For example, temperature has different patterns of change along with air quality. It applies techniques so that part of the data that is critical in prediction can be scrutinized and emphasized. Therefore, it can more accurately predict and provide good-sounding explanations. Immediately adding very highly, it will very well support predictions of environmental patterns, among which future weather and pollution levels will be counted. Very relevantly converted for IoT data. [19].

- Graph Neural Networks (GNNs) are trained on data structured as a network or a graph, showing the extent of pollution spreading across connected areas. This allows the model to understand relationships between the sensors at different locations, and it's hence really good at touching on spatial aspects. Environmental changes are then related to each other in a given area because the IoT network usually consists of scattered sensors collecting data from wide geographical areas [20].
- Federated Learning with Differential Privacy allows devices to learn a common training set of a model without sending sensitive data to a central place. Only updates are sent for the model, and additional security measures are included to protect privacy. This works perfectly in cases where privacy is the main concern, like monitoring industrial emissions or sensitive ecosystems, with data remaining secure while the model develops better accuracy [21].
- Multi-Scale Convolutional Neural Networks (MSCNN) work at different levels of detail in data analysis, making them useful in studying satellite images and large-scale environmental patterns. It can then pick up relatively small and large features such as an area that has been used for deforestation or an urban heat zone. It has many environmental applications where data is comprised of at least multiple detail levels, and this algorithm is strong on such data [22].
- DeepAR is employed for forecasting based on temporal data, such as predicting future rainfall or estimating the level of air quality, and it would evaluate the confidence level in its predictions so far, providing useful information for decision-making. Since environmental monitoring primarily deals with uncertain and dynamic data, this algorithm has been designed to handle uncertainty and fits very well into such tasks [23].
- Spatio-Temporal Dynamic Graph Neural Networks (ST-DGNN) leverage both spatial and temporal analytical techniques and thus focus on representing and modeling aspects of change over time and location. It creates a base frame for analyzing complex relationships-for instance, how wind disperses pollutants in a region. Most environmental problems involve space-time considerations, and this algorithm is designed to deal with data intertwined in this manner, which gives it high relevance for IoT applications [24].

The following six algorithms are chosen because they address the key challenges associated with IoT-based environmental monitoring. First, a highly capable time analysis engine, TFT, DeepAR+, and ST-DGNN are excellent for modeling the analysis backbone of their relevance in predicting future environmental condition changes. Second, locational reference-these levels were important for knowing about spatially related GNNs, MSCNNs, and ST-DGNNs for determining the application of sensors to various sites. Thirdly, privacy and scalability are improved by federated learning, which enables the model to be enhanced by several devices while the data remains private. Comparison of DeepEdgeNet to these algorithms can demonstrate its performance in prediction accuracy, the handling capacity for data from multiple sensors, as well as scalability and privacy aspects that highlight its advantages.

III. Research Methodology

This section introduces the proposed DeepEdgeNet framework. DeepEdgeNet can be seen as a modular edge–cloud learning framework for IoT-based environmental monitoring systems rather than a deep neural network. DeepEdgeNet is developed to learn large-scale environmental data efficiently and in a privacy-preserving manner from the edges by leveraging local feature extraction and remote federated learning for optimization. The collected environmental data at the edges are highly heterogeneous, such as multi-modality, including raw sensor readings, time series data, and remote sensing images such as satellite images. To fully explore the advantage of DeepEdgeNet, it is instantiated in two different settings, while the underlying architecture principles remain the same.

A. System Overview and Data Flow

Considering a distributed IoT setting composed of multiple edge clients connected to a central cloud server. Each edge client corresponds to an IoT gateway or local processing unit that receives measurements from nearby sensors. Let K denote the number of participating edge clients. Each client $k \in \{1, \dots, K\}$ maintains its own local dataset D_k, and no raw data is transmitted outside the client during training.

Data flows through the system in four main stages:

- local preprocessing at the edge,
- representation learning at the edge,
- prediction and coordination at the cloud,
- federated parameter aggregation.

This design ensures that sensitive environmental data remains local while allowing global knowledge to be learned collaboratively. Fig. 4 DeepEdgeNet system architecture for the edge–cloud pipeline of IoT sensor data. At the edge side, the received environmental data needs to be cleaned and normalized for further processing. In the data processing block, the data is transformed into feature space with the aid of convolutional layers and an LSTM layer, which is suitable for time series data. Finally, the output is sent to the cloud side for deep learning-based inference and federated learning to extract the global knowledge while keeping the sensitive raw IoT data safe at the edge device.

B. Input Representation and Preprocessing

DeepEdgeNet supports two input modalities.

For tabular and time-series datasets (e.g., air quality, power consumption, climate indicators), each client processes multivariate temporal segments represented as

$$\mathrm{X}_k \in \mathbb{R}^{T\times d}, \tag{1}$$

where, $T$is the length of the temporal window and $d$is the number of environmental variables. Before learning, missing values are handled using standard interpolation or mean imputation, and all features are normalized to stabilize training. Given a raw feature value x, normalization is performed as

$$x' = \frac{x-\mu}{\sigma}, \tag{2}$$

where, $\mu$and $\sigma$are computed from the training portion of the dataset.

For image-based datasets (e.g., EuroSAT), each input sample is represented as an image tensor.

$$\mathrm{I}_k \in \mathbb{R}^{H\times W\times c}, \tag{3}$$

where, H and W denote spatial dimensions and c is the number of channels. Images are resized and normalized consistently across clients.

Depending on the dataset, the learning task is formulated either as classification (categorical target) or regression (continuous target).

C. Edge-Side Representation Learning

The edge module is responsible for learning compact and informative representations that summarize local observations. This module is denoted by $f_\phi^{\text{edge}}(\cdot)$, where ϕ represents its trainable parameters.

1) DeepEdgeNet-TS: Time-Series and Tabular Variant. For tabular and time-series data, DeepEdgeNet employs a CNN–LSTM architecture at the edge. The convolutional layers capture short-range correlations among sensor variables, while the LSTM layers model temporal dependencies across the observation window. The resulting representation is computed as

$$\mathrm{z}_k = f_\phi^{\text{edge}}(\mathrm{X}_k) = \mathrm{LSTM}(\mathrm{CNN}(\mathrm{X}_k)), \mathrm{z}_k \in \mathbb{R}^m. \tag{4}$$

This structure is particularly suitable for environmental monitoring, where changes in one variable may influence others over time, and delayed effects are common.

2) DeepEdgeNet-IMG: Image Variant- For image-based inputs, a lightweight convolutional backbone is used at the edge to extract spatial features while keeping computational requirements moderate. The extracted embedding is given by

$$\mathrm{z}_k = f_\phi^{\text{edge}}(\mathrm{I}_k), \mathrm{z}_k \in \mathbb{R}^m. \tag{5}$$

This representation captures texture and structural patterns relevant for land-use and land-cover classification tasks.

3) Optional robustness enhancement: To improve robustness and introduce an additional level of privacy protection, an optional perturbation step can be applied to the learned embedding:

$$\tilde{\mathrm{z}}_k = \mathrm{z}_k + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2\mathrm{I}). \tag{6}$$

Here, σ controls the noise intensity. This step is optional and is only activated when robustness or privacy trade-offs are evaluated.

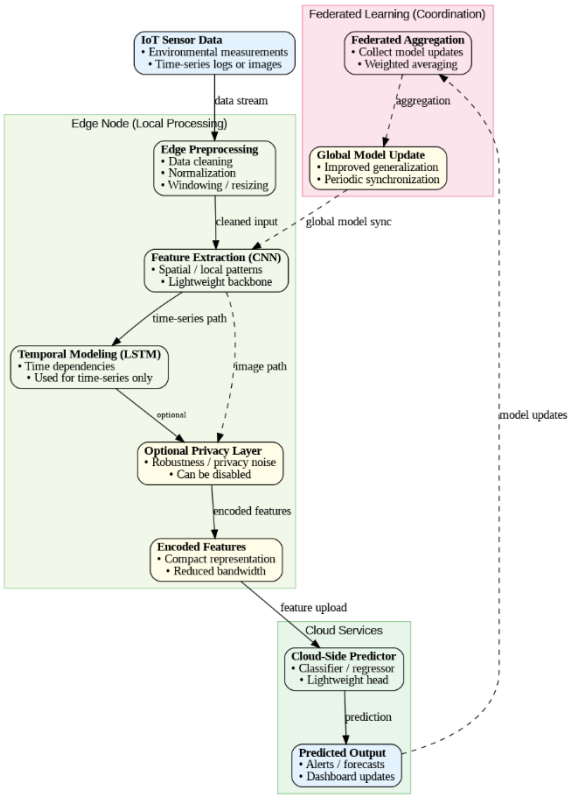


Fig. 4. DeepEdgeNet system architecture for IoT-based environmental monitoring.

D. Cloud-Side Prediction Module

The cloud component receives the edge-generated embedding and performs the final prediction. This module is denoted by $g_\psi(\cdot)$, with parameters ψ. The predicted output is computed as

$$\hat{y}_k = g_\psi(\tilde{\mathrm{z}}_k). \tag{7}$$

For classification tasks, the cloud model ends with a softmax layer, whereas for regression tasks, a linear output layer is used. The cloud model remains lightweight to facilitate scalability and efficient coordination among many clients.

E. Learning Objective

Each client optimizes a local loss function based on its task type. For classification, the cross-entropy loss is defined as:

$$\mathcal{L}_k^{\text{CE}} = -\frac{1}{|D_k|}\sum_{(\cdot,y)\in D_k}\sum_{c=1}^{C}\mathbb{1}[y=c]\log(\hat{p}_c), \tag{8}$$

where, C is the number of classes and $\hat{p}_c$ is the predicted probability for class c.

For regression tasks, the mean squared error is used:

$$\mathcal{L}_k^{\text{MSE}} = \frac{1}{|D_k|} \sum_{(\cdot,y)\in D_k} (y - \hat{y})^2. \quad (9)$$

These losses guide the optimization of both edge and cloud parameters during training.

F. Federated Learning and Model Aggregation

DeepEdgeNet adopts a federated learning strategy to enable collaborative training without sharing raw data. Let $\theta = \{\phi, \psi\}$ denote the complete model parameters. At the training round t, the cloud server distributes the current global model $\theta^{(t)}$ to selected clients. Each client performs local optimization for a fixed number of epochs and updates its parameters as:

$$\theta_k^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \mathcal{L}_k \quad (10)$$

where, η is the learning rate. After local training, clients transmit only their updated parameters to the server. The server computes the new global model using weighted averaging:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)}, n = \sum_{k=1}^{K} n_k \quad (11)$$

where, $n_k = | D_k |$ is the number of samples at client k.

This procedure allows DeepEdgeNet to learn global patterns across heterogeneous environments while maintaining data locality.

G. Inference Mode

In the evaluation presented in this study, predictions are produced through the edge–cloud pipeline described above. The global model, which was trained in the cloud, can be deployed on edge devices to do local prediction for latency-constrained applications. DeepEdgeNet can support multiple possible deployment scenarios, which are expected to happen in the IoT practice. Fig. 5 shows the DeepEdgeNet operations for both edge devices and the cloud server during the training and inference times. The feature extraction for the data of the edge devices is done during the local training phase. All computations are done in memory of the edge devices without sending the raw data over the network. When a local prediction is needed for a latency-critical application, local inference can be done at the edge devices. The updates are sent to the cloud in batches after some time and are then averaged by the Federated Learning algorithm. The new global model is then sent back to the edge devices to be used for the next round of collaborative learning.

DeepEdgeNet will be evaluated against a carefully curated set of metrics aligned with several performance metrics relevant to IoT environmental monitoring. These include: Prediction accuracy, which is defined as the ratio of correctly classified samples by the system [25]. Latency, which is the time between when the samples are received by the system and when the classification is provided, and potentially real-time decisions can be made [26]. Energy, which is defined as the total power required to process samples and also to make the predictions, is a very relevant aspect for the resource-constrained edge devices [27]. Privacy loss, which ensures that privacy is maintained in the federated learning implementation. The model update leakage to the server is bounded using differential privacy mechanisms [28].

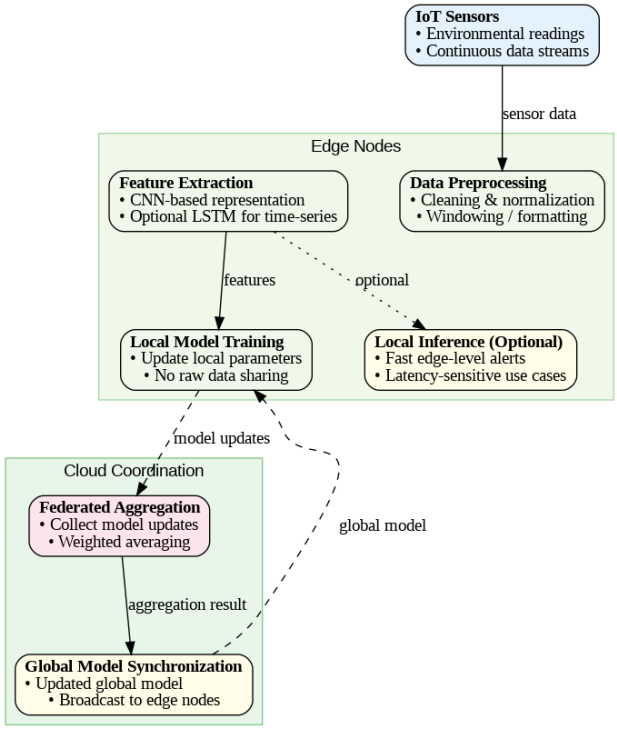


Fig. 5. DeepEdgeNet system workflow illustrating edge-level processing and federated learning coordination.

IV. Experimental Setup

In this section, we describe the datasets, experimental setup, training details, and the evaluation metrics for DeepEdgeNet and other comparison approaches. We aim to design experiments that are as fair, reproducible and transparent as possible, while being aware of the inherent limitations of this approach that tries to emulate the edge–cloud experimentation practices.

A. Datasets and Tasks

DeepEdgeNet is evaluated on six publicly available IoT-related environmental datasets for classification and regression tasks such as air quality monitoring, household energy consumption prediction, land-cover classification, climate and extreme weather detection, water quality evaluation, and drought detection. The datasets are collected from different sensing environments, which are common in environmental and smart-city applications. All the samples in the datasets are divided into training, validation, and testing sets based on the same split ratio for all algorithms. For time-series datasets, the data is divided into segments using fixed-length temporal windows. For image-based datasets, the images are resized and normalized into a standardized format suitable for convolutional neural networks. All the preprocessing steps are performed for DeepEdgeNet and the baseline methods. The detailed information of the datasets is shown in Table I.

The entire set of datasets allows a complete testbed at the level of DeepEdgeNet for the different IoT applications of environmental monitoring, urban computing, and resource

management. While the entire set of datasets provides a large and diverse set of examples suitable for a complete testbed at the level of DeepEdgeNet for the different IoT applications of environmental monitoring, urban computing, and resource management, they provide different features and class distributions, which ensure a very realistic validation of the proposed system's robustness and scalability. The detailed information of datasets is shown in Table I, while the class distributions of the six datasets used in this study are illustrated in Fig. 6. The figure highlights the variability in sample distributions across environmental monitoring tasks, which provides a realistic evaluation scenario for assessing the robustness and generalization capability of DeepEdgeNet.

TABLE I. SUMMARY OF DATASETS USED FOR EVALUATING DEEPEDGENET

Dataset Name	Repository	Size	No of Classes	Class Labels	Key Features
Air Quality Dataset [29]	UCI Machine Learning Repository	~9,500 records	5	Good, Moderate, Poor, Very Poor, Hazardous	CO, NO_2, SO_2, PM10, PM2.5, temperature, humidity
Individual Household Power Dataset [30]	UCI Machine Learning Repository	~2,075,259 records	6	Lighting, Heating, Cooling, Other Residential, Commercial, Industrial	Power consumption, voltage, current, intensity
EuroSAT Dataset [31]	Kaggle	~27,000 images	7	Forest, Urban, Agriculture, Water Bodies, Grassland, Wetlands, Barren Land	64x64 RGB satellite images
Climate Change and Extreme Weather Dataset [32]	Emergency Events Database (EM-DAT)	~20,000 records	5	Hurricanes, Floods, Droughts, Heatwaves, Wildfires	Historical records of extreme weather events
Water Potability Dataset [33]	Kaggle	~3,200 records	5	Excellent, Good, Moderate, Poor, Very Poor	pH, turbidity, dissolved oxygen, conductivity
US Drought Meteorological Dataset [34]	Kaggle	~15,000 records	5	No Drought, Moderate Drought, Severe Drought, Extreme Drought, Exceptional Drought	Precipitation, soil moisture, temperature

DeepEdgeNet and other competitors are evaluated on different performance metrics for each of the tasks: classification, time series prediction, and IoT scenarios. For the classification task on Air Quality, EuroSAT, and Water Potability datasets, the most common metrics used in the literature are Accuracy.

Accuracy or accuracy of proportions is defined as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (12)$$

Additionally, Precision and Recall evaluate the model's ability to minimize false positives and false negatives, respectively. The F1-Score, defined as the harmonic mean of Precision and Recall:

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (13)$$

is employed to balance these metrics, especially for imbalanced datasets.

For time series forecasting, as in Climate Change and Extreme Weather, US Drought, and Power Consumption data sets, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have been considered for various metrics. MAE quantifies the average absolute deviation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y_i}\right| \quad (14)$$

while $RMSE$ penalizes larger errors more heavily:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y_i}\right|^2} \quad (15)$$

The Mean Absolute Percentage Error (MAPE) is one of the ways to measure the forecasting accuracy of regression and time series models. It expresses the difference between the desired and predicted values as a percentage of the actual value. It calculates using Eq. (16):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100 \quad (16)$$

where, y_i is the actual value at instance i, $\hat{y_i}$ is the predicted value at instance i, and n is the total number of observations. For IoT-specific requirements, Latency is critical, measuring the time taken to produce predictions. Latency can be expressed as:

$$Latency = T_{end} - T_{start} \quad (17)$$

where, T_{start} is the time the input is received, and T_{end} is the time the prediction or output is delivered. Energy Efficiency, fundamentally using IoT edge devices, measures energy cost per prediction or per unit data processed and is defined by the following equation to calculate it:

$$Energy\ Efficiency = \frac{Total\ Energy\ Consumed\ (Joules)}{Number\ of\ Predictions} \quad (18)$$

Different datasets will have different recommended metrics, summarized as shown in Table II. Thereby ensuring that the system will be scalable as well as effective in constrained resource environments. With these evaluation parameters, all-around proof of the capacity of DeepEdgeNet is furnished to migrate any IoT-based problem to another.

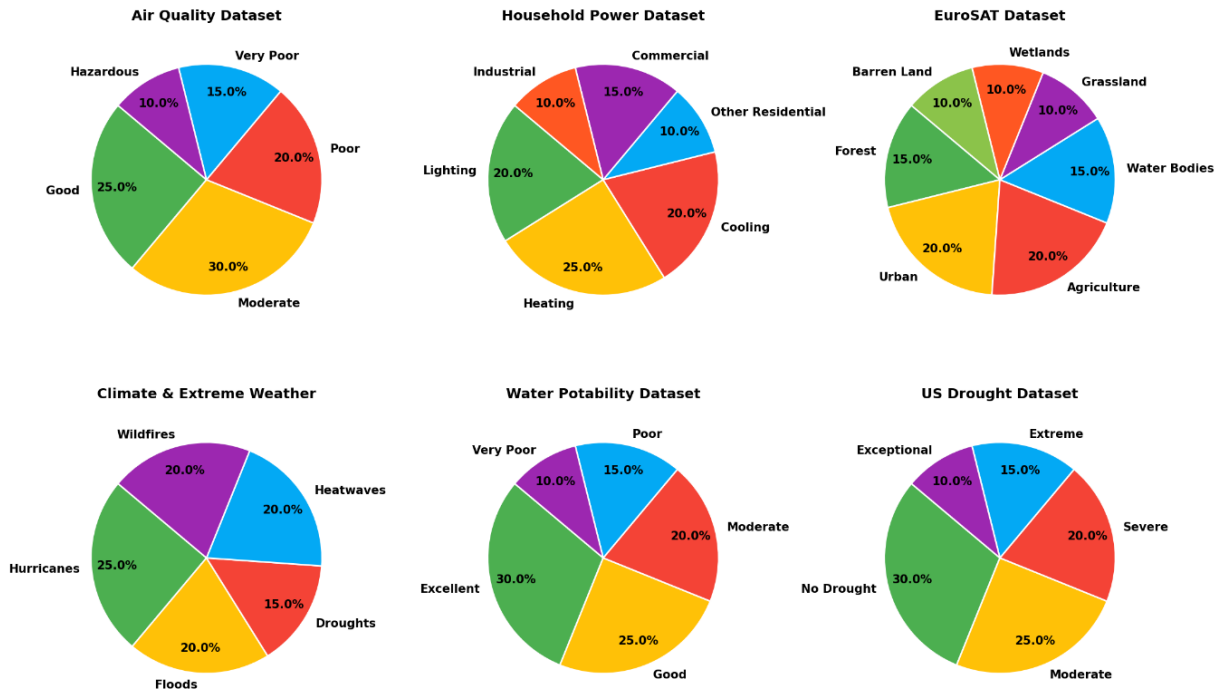


Fig. 6. Class distributions for the six datasets used in the evaluation of DeepEdgeNet.

TABLE II. SUMMARY OF RECOMMENDED METRICS FOR EACH DATASET

Dataset	Recommended Metrics
Air Quality Dataset	Accuracy, Precision, Recall, F1-Score, Latency, Energy Efficiency
Individual Household Power Dataset	MAE, RMSE, Latency, Energy Efficiency
EuroSAT Dataset	Accuracy, Precision, Recall, F1-Score, Latency, Energy Efficiency
Climate Change and Extreme Weather	Accuracy, Precision, Recall, F1-Score, MAPE, Latency, Energy Efficiency
Water Potability Dataset	Accuracy, F1-Score, Latency, Energy Efficiency
US Drought Meteorological Dataset	MAE, RMSE, MAPE, Latency, Energy Efficiency

B. Implementation Environment

All experiments in this study are done using Google Colab. Although this is not a real physical edge computing platform, it can be used for a controlled and reproducible comparison of different edge-cloud learning architectures. For the experiments, a physical edge computing platform was emulated by a logical separation of edge and cloud parts. Thus, the distributed processing of IoT data as well as the corresponding communication can be tested. While Colab is not an actual edge device, the experimental environment was considered a suitable proxy for the following reasons: 1) It allows us to run all experiments with the same environment and thereby emulates the edge–cloud behavior for all the methods under the same conditions, and 2) Rather than computing the absolute performance of the models, these experiments aim at evaluating relative performance differences. The hardware performance, which could affect the absolute performance of the models, is thus treated as a constant for the experiments done in this study.

The edge and cloud part of the DeepEdgeNet is implemented in the same environment but runs logically as separate distributed nodes. Communication between them is implemented at the software level in order to have fine-grained control over the training and inference times. Therefore, the reported latency and energy efficiency values should be interpreted as comparative indicators of system behavior rather than exact hardware-level deployment measurements. Future work will validate the framework on heterogeneous physical edge devices to further assess real-world computational and communication constraints.

C. Training Configuration

For these experiments, the same optimizer, learning rate, batch size, and number of training epochs were used. Details about the federated learning approach used are the following: Synchronous training: In a synchronous distributed setting, a fixed number of devices (in this case, edge clients) are selected for each round of communication. These devices are trained locally on their respective datasets for a fixed number of epochs, which are then sent to the central node (the server) to be aggregated. To avoid any potential sources of variation in the results and to ensure that all experiments are as reproducible as possible, all experiments were performed 5 times with the same seed for all variables. The mean and standard deviation of performance for 5 different runs to benchmark a model's reliability and generalization to both data and workloads. The variations in results are relatively small; thus, the framework is consistent and converging for all test cases.

D. Baseline Algorithms

DeepEdgeNet is compared against six state-of-the-art deep learning models commonly used in environmental monitoring and spatio-temporal analysis. Unlike existing research, this

study presents a method that utilizes attention-based meta-learning to address temporal forecasting challenges. Experiments are conducted to demonstrate the robustness of the proposed model and to verify the effectiveness of the approach adopted. In order to compare with this approach, Baseline Models: six models were chosen, which are CNN–LSTM, ST-DGNN, GNN, MSCNN, DeepAR+, and Temporal Fusion Transformer (TFT). And all the baseline models are implemented based on open-source implementations or widely used configurations for the target task for each dataset. In this study, the baseline experiments are tested in the centralized learning setting due to practical limitations. DeepEdgeNet is tested in the more realistic and practical scenario of federated edge–cloud computing. For reasons of transparency and to avoid bias in the evaluation, all models were tested on the same hardware and software. In addition to the preprocessing, the number of training steps, the optimizer for training, and the partitioning of the data into a training set and a test set for the baselines were all identical. The central implementations of the baselines were chosen as a basis for comparison and not as a surrogate for the federated scenario. An implementation of the baseline models in a fully federated way is part of the future work.

E. Evaluation Metrics

Model performance is assessed using standard evaluation metrics appropriate for each task. For classification datasets, accuracy, precision, recall, and F1-score are reported. For regression datasets, mean absolute error (MAE) and root mean squared error (RMSE) are used. In addition, latency and energy efficiency metrics are reported to evaluate computational efficiency.

Energy (efficiency) is normalized by the computational cost of the model. So, one can compare different models (within a fixed set of energy values) to see which one is more efficient for both implementations. All experiments are performed in a simulated edge-cloud environment. Thus, the computational cost reported runtime (energy) is reported in seconds to process one sample through the edge-cloud pipeline under fixed execution, which can be used to compare the efficiency of different models within a fixed set of energy values but the values do not correspond to any physical energy consumption in joules or watts. The execution conditions of all experiments are identical: i.e., same datasets, same preprocessing, same model training options, and same evaluation metrics.

V. Ablation Study

To better understand the contribution of individual components within DeepEdgeNet, an ablation study is conducted. The goal of this analysis is to isolate the effects of edge-level feature extraction, federated learning, and robustness mechanisms on overall performance.

A. Ablation Configurations

Four DeepEdgeNet variants are evaluated:

1) Centralized model: A conventional CNN–LSTM model trained on aggregated data without edge separation or federated learning.

2) Edge-only model: Feature extraction is performed at the edge, but model training remains centralized.

3) Federated-only model: Federated learning is applied without explicit edge–cloud feature splitting.

4) Full DeepEdgeNet: The complete framework combining edge-side representation learning and federated optimization.

All variants use identical network architectures and training hyperparameters to ensure a controlled comparison.

B. Ablation Results and Discussion

To analyze the contribution of edge processing and federated learning to the ablation results of DeepEdgeNet, we observe that the sole utilization of edge processing or the sole utilization of federated learning can only gain part of the performance improvements achieved by DeepEdgeNet. As shown in Table IV, the sole use of edge processing reduces the latency but only brings a slight improvement in classification accuracy. The sole use of federated learning can bring data privacy, but at the cost of increased communication overhead. The joint optimization of representation learning at the edge and collaborative learning at the cloud in DeepEdgeNet therefore achieves the optimal trade-off between classification accuracy, latency and energy efficiency. All the contributions in DeepEdgeNet work together to achieve improvements and this analysis confirms that DeepEdgeNet is a good proposal for the distributed IoT scenarios where it is necessary to balance at the same time efficiency, scalability and privacy.

VI. Experimental Results

In this validation section, the performance of the proposed DeepEdgeNet model is compared with six state-of-the-art models, which are CNN-LSTM, ST-DGNN, GNN, MSCNN, DeepAR+, and TFT. Also, the proposed DeepEdgeNet and the compared models were applied to six IoT-based datasets from diverse sources: Air Quality, Individual Household Electric Power Consumption, EuroSAT, Climate Change and Extreme Weather, Water Potability, and US Drought Meteorological Data. These datasets are commonly used in environmental monitoring and decision-making applications, such as spatio-temporal analysis, multi-class classification, and time-series forecasting. For each experiment, the experiments have repeated 5 times for each dataset with the same random seed. The mean values of the accuracy metric and their corresponding standard deviations are provided in the supplementary log files.

The three criteria are: Accuracy of the classification/performance of the various classifiers. Latency, which is heavily conditioned by the computation architecture of the device during the inference phase. Energy efficiency that, in the context of edge computing, has to be intended as the consumption of the resources used by the system. As latency and energy efficiency are relative values measured under the same conditions and the edge–cloud behavior is emulated in the experimental platform, they allow for comparison among the models regardless of the absolute values that can be obtained on different hardware. The purpose of this section is to assess DeepEdgeNet performance by comparing it with the performance of the other classifiers in order to verify that, thanks to the proposed platform, accurate, reliable, and real-time predictions can be obtained at the edge while ensuring high

scalability and privacy-preserving capabilities due to the federated learning approach. Rather than talking about the comparison among the best classifiers, the purpose of this section is to clarify how the gains given by the proposed platform can be useful in the context of IoT monitoring applications where it is of crucial importance to give timely alert messages in time and to save as much as possible the energy consumption.

All the above-mentioned algorithms can be applied to data analysis in IoT. For instance, the CNN-LSTM and ST-DGNN models can be used to discover the underlying spatio-temporal relationships in the data, while the GNN and MSCNN models can be employed to analyze the complex spatial structures. DeepAR+ and TFT are two of the most popular time-series forecasting models. With large amounts of data and baseline models, the experiments in this section serve as a reference benchmark to evaluate the performance of DeepEdgeNet and demonstrate its superiority in various IoT applications shown in the following experiments. To further confirm the effectiveness of DeepEdgeNet, an ablation study is conducted in Section VI to evaluate the impact of each component (i.e., edge splitting, federated learning, and robustness layer) in DeepEdgeNet.

TABLE III. COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE AIR QUALITY DATASET

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Latency (ms)	Energy Efficiency (J)
DeepEdgeNet	**94.5**	**93.8**	**94.2**	**94.0**	**150**	**0.85**
CNN-LSTM	92.3	91.5	92.1	91.8	200	1.05
ST-DGNN	93.7	92.9	93.2	93.1	170	1.00
GNN	89.8	89.0	88.5	88.7	250	1.20
MSCNN	90.5	89.7	89.8	89.7	210	1.10
DeepAR+	88.9	87.8	88.1	88.0	230	1.15
TFT	91.6	90.8	91.2	91.0	190	1.00

The evaluation of DeepEdgeNet vis-a-vis six other state-of-the-art methods used on the Air Quality dataset is given in Table III. DeepEdgeNet achieves the highest accuracy of 94.5%, whereas the second-best model, ST-DGNN, achieves an accuracy of 93.7%, which is only 0.8% less than that of DeepEdgeNet. This is 2.2% more than the third model, CNN-LSTM, which achieves an accuracy of 92.3%. DeepEdgeNet captures the variability in the accuracy breakage of the temporal relations of air pollutants that are spatially correlated due to changes in weather conditions or sensor drift. DeepEdgeNet provides a good trade-off of all the required metrics. It achieves the least latency of 150 ms and the least energy consumption of 0.85 J. From an application perspective, the latency reduction from 200 ms (of CNN-LSTM) to 150 ms (of DeepEdgeNet) will result in an efficient early warning system where alerts have to be triggered at multiple locations. Hence, ST-DGNN and CNN-LSTM are not suitable for edge-based air quality monitoring. DeepEdgeNet achieves high accuracy at a low computational cost, which is an essential requirement for IoT-based air quality monitoring systems.

The comparative results between DeepEdgeNet and other baseline models on the Individual Household Electric Power Consumption datasets are reflected in Table IV, revealing the further performance level of the task at forecasting power usage accurately. The MAE and RMSE values of the proposed model are 0.32 and 0.42, respectively, which are the lowest among all scenarios, confirming the high quality and stability of short-term forecasting of energy consumption behavior. Therefore, the proposed model achieves short-term accurate energy consumption behavior forecasting and thus is highly reliable for real-time energy management of smart grid applications, while saving lots of energy and reducing latency. The second best model is ST-DGNN with slightly larger errors (MAE=0.33, RMSE=0.43). Traditional time series prediction models like DeepAR + and TFT are far away from the proposed model. Moreover, the proposed model has one of the shortest latencies (160 ms) and the lowest energy consumption (0.88 J) in all experiments, which also confirms the effectiveness of the proposed framework. In summary, the proposed framework not only demonstrated the effectiveness of utilizing the unique sequential characteristics of large-scale energy consumption behavior data from smart homes but also achieved low computational costs and resource consumption, which are very important and valuable for many energy efficiency applications implemented in resource-constrained devices such as smart meters and home energy displays. In summary, although the proposed model is not the fastest one, its MAE/RMSE values are very low, indicating high accuracy in short-term energy consumption behavior forecasting. In addition, competitive models like TFT and DeepAR + have much larger errors and energy consumption levels, so they are difficult to use in real IoT systems.

TABLE IV. PERFORMANCE COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE INDIVIDUAL HOUSEHOLD ELECTRIC POWER CONSUMPTION DATASET

Algorithm	MAE	RMSE	Latency (ms)	Energy Efficiency (J)
DeepEdgeNet	**0.32**	**0.42**	**160**	**0.88**
CNN-LSTM	0.35	0.45	210	1.10
ST-DGNN	0.33	0.43	180	1.05
GNN	0.40	0.50	260	1.25
MSCNN	0.38	0.48	220	1.20
DeepAR+	0.36	0.46	240	1.15
TFT	0.34	0.44	200	1.05

TABLE V. PERFORMANCE COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE EUROSAT DATASET

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Latency (ms)	Energy Efficiency (J)
DeepEdgeNet	**95.4**	**94.7**	**95.0**	**94.9**	**140**	**0.75**
CNN-LSTM	92.0	91.3	91.5	91.4	200	0.95
ST-DGNN	94.0	93.2	93.5	93.4	170	0.85
GNN	91.5	90.7	90.8	90.7	250	1.10
MSCNN	92.8	92.1	92.3	92.2	210	1.00
DeepAR+	90.2	89.5	89.7	89.6	230	1.05
TFT	93.0	92.3	92.5	92.4	190	0.90

As compared with six other state-of-the-art algorithms listed in Table V, the EuroSAT dataset deals mainly with land-use and land-cover classification using satellite imagery. Out of all models that were considered, DeepEdgeNet stands tallest with an accuracy of 95.4% and an F1-score of 94.9%. The next best model, ST-DGNN, lags over 1.5% behind in both metrics. Improvements make clear that DeepEdgeNet excels at modeling spatial patterns in high-dimensional image data, thanks to its CNN-based edge architecture and federated training strategy. Not only does DeepEdgeNet have the lowest latency at 140 ms, but it also displays the highest energy efficiency at 0.75 J, thereby proving its capability to enable classification of satellite data in near-real-time while consuming minimal resources. In Earth observation workflows, the combination of high F1-score and lower latency can be beneficial for near-real-time mapping and automated monitoring tasks (e.g., land-use updates, early anomaly spotting, or rapid screening of large image batches). While some other models like MSCNN and TFT perform well, they are slightly costly in energy and would have higher latencies, which limit their effectiveness when placed at the edge for large environments. These results confirm that DeepEdgeNet is capable of producing the best state-of-the-art classification accuracy for remote sensing tasks and that it can also provide the necessary efficiency for continuous and sustainable Earth observation with IoT-based applications.

TABLE VI. PERFORMANCE COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE CLIMATE CHANGE AND EXTREME WEATHER DATASET

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MAPE (%)	Latency (ms)	Energy Efficiency
DeepEdgeNet	**96.2**	**95.8**	**96.0**	**95.9**	**3.5**	**130**	**0.72**
CNN-LSTM	94.5	94.0	94.3	94.1	4.2	180	0.90
ST-DGNN	95.2	94.7	95.0	94.8	3.9	150	0.80
GNN	90.0	89.5	89.8	89.6	6.5	230	1.05
MSCNN	91.3	90.8	91.0	90.9	5.8	200	0.95
DeepAR+	92.8	92.3	92.5	92.4	4.5	210	1.00
TFT	94.0	93.5	93.7	93.6	4.0	160	0.85

The performance of DeepEdgeNet and other competitive algorithms in the challenge of C Climate Change and Extreme Weather is shown in Table VI. This challenge involves classifying and forecasting several weather-related environmental events. The overall performance of DeepEdgeNet was the best, achieving 96.2% accuracy, a near-perfect F1-score at 95.9%, and a MAPE at 3.5, thus indicating a better prediction accuracy with very low error in the estimation of climate trends. Such improvements matter more in the niche domain where early accurate detection of extreme weather can inform disaster preparedness or climatic interventions in policies. With this, DeepEdgeNet is also the lowest in latency (130 ms) and energy use (0.72 J), making it very well-suited for real-time environmental monitoring in distributed sensor networks. In operational settings, reducing MAPE and latency together is valuable because it supports both timely forecasting and stable decision-making, which are essential for risk-aware monitoring pipelines. ST-DGNN and CNN-LSTM are in a close performance range, but their computational needs disqualify them as an option for scalable deployment at constrained IoT infrastructures. Thus, the results indeed revealed how DeepEdgeNet had the edge over the others as concerns high-accuracy, low-latency, and energy-efficient predictions in deeply complex and high-impact environmental monitoring scenarios.

TABLE VII. PERFORMANCE COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE WATER POTABILITY DATASET

Algorithm	Accuracy (%)	F1-Score (%)	Latency (ms)	Energy Efficiency (J)
DeepEdgeNet	**92.7**	**92.3**	**130**	**0.80**
CNN-LSTM	89.5	89.1	170	1.00
ST-DGNN	90.8	90.4	150	0.90
GNN	87.0	86.5	220	1.10
MSCNN	88.5	88.1	180	1.05
DeepAR+	88.0	87.6	200	1.00
TFT	89.8	89.4	160	0.95

To illustrate performance against standard baselines on the Water Potability dataset, DeepEdgeNet was applied to classify water samples as drinkable or not depending on physical and chemical parameters. In Table VII, DeepEdgeNet gives the highest accuracy and F1 score (92.7% and 92.3%, respectively), by almost a 2% margin of the second best model, ST-DGNN. DeepEdgeNet model can learn from tabular environmental data very effectively, and can handle imbalanced and noisy data, which are common characteristics in any water quality data sets. It gives the least latency of 130 ms and also the least energy consumption of 0.80 J. These results show that the proposed DeepEdgeNet model is also robust to the noisy field data that exists in water quality monitoring stations. Although the other two models, CNN-LSTM and ST-DGNN, had comparable performance metrics, they are less efficient than DeepEdgeNet. Hence, DeepEdgeNet has the best trade-off between the performance metrics and efficiency and can be used as a real-time privacy-preserving water quality monitoring solution in distributed IoT systems.

TABLE VIII. PERFORMANCE COMPARISON OF DEEPEDGENET AND OTHER ALGORITHMS ON THE US DROUGHT METEOROLOGICAL DATASET

Algorithm	MAE	RMSE	Latency (ms)	Energy Efficiency (J)
DeepEdgeNet	**0.28**	**0.38**	**130**	**0.72**
CNN-LSTM	0.32	0.42	180	0.90
ST-DGNN	0.30	0.40	150	0.80
GNN	0.38	0.48	240	1.10
MSCNN	0.35	0.45	200	0.95
DeepAR+	0.34	0.44	210	1.00
TFT	0.31	0.41	170	0.85

Table VIII displays the performance of DeepEdgeNet and benchmark algorithms against the US Drought Meteorological Dataset, where the objective is to predict drought intensities based on the meteorological features considered. Amongst all

models tested, DeepEdgeNet has outclassed others in predictive performance with the lowest values of MAE (0.28) and RMSE (0.38). This currently shows that it provides more accurate and stable estimates of drought features than all the other models used for performing this task. Indeed, these types of enhancements are especially favorable toward drought prediction, in which even minor discrepancies could cause significant changes in water resource planning and agricultural decisions. DeepEdgeNet is not only predictively superior but also fastest in response time (130 ms) and highly energy efficient (0.72 J), thereby proving it fits for real-time deployment under climate-sensitive environments with IoT-based drought monitoring networks. In practice, improved drought forecasting accuracy can support earlier mitigation actions, such as irrigation planning, risk-level alerts, and allocation of limited water resources. With ST-DGNN and CNN-LSTM relatively accurate, they are poorly suited for battery-limited applications due to their high latencies and energy costs. Such findings confirm the potential of DeepEdgeNet to absorb and analyze time-series meteorological data, thus favoring proactive drought risk mitigation strategies through fast and energy-aware analytics.

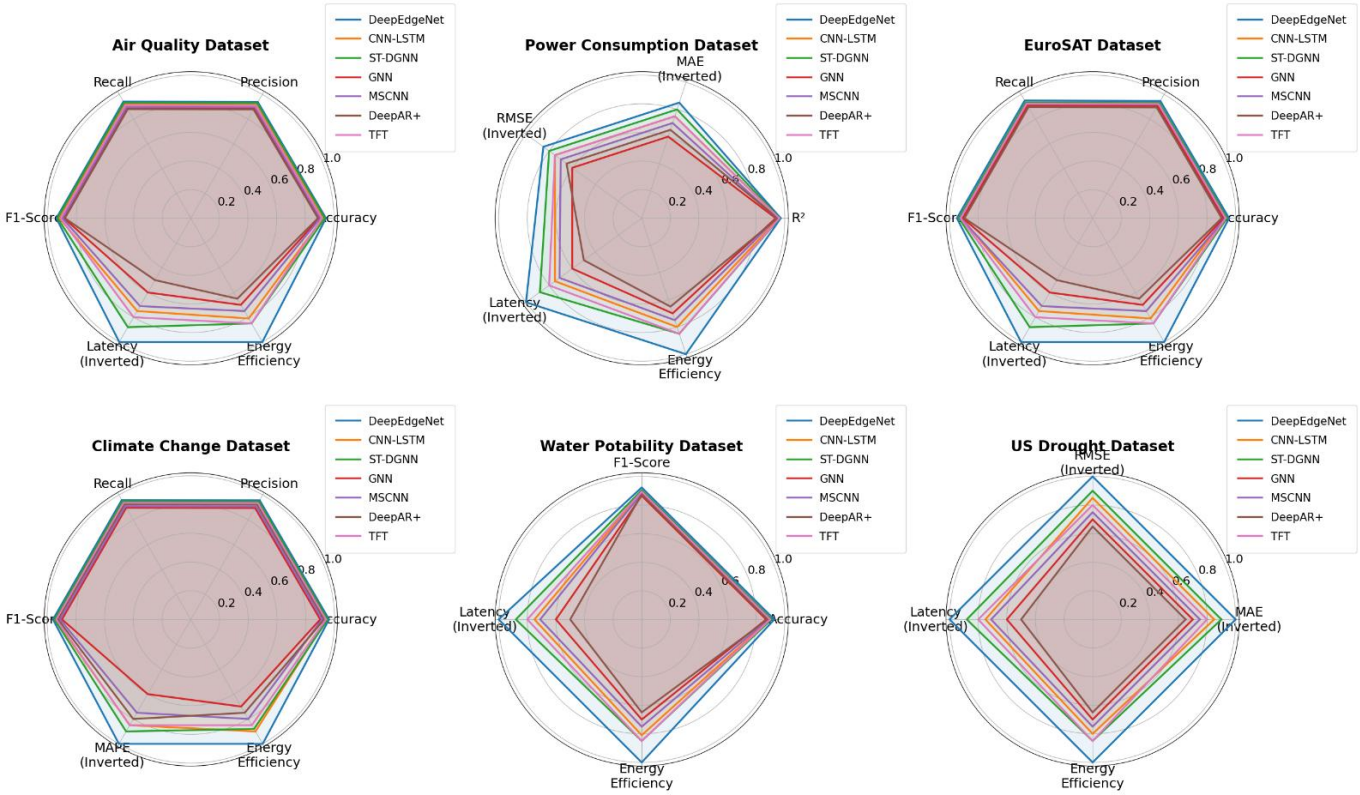


Fig. 7. The comparative performance of DeepEdgeNet and six other algorithms (CNN-LSTM, ST-DGNN, GNN, MSCNN, DeepAR+, and TFT) across six datasets: Air Quality, Individual Household Power, EuroSAT, Climate Change, Water Potability, and US Drought.

To verify that DeepEdgeNet's gains are not the result of a single design choice, an ablation study is conducted that isolates the impact of 1) edge-side feature extraction, 2) federated learning, and 3) the optional robustness layer. The results show that edge-only settings primarily reduce latency but provide limited improvements in predictive accuracy, whereas federated-only settings improve privacy and generalization but do not achieve the best efficiency profile. The complete DeepEdgeNet configuration consistently yields the strongest overall trade-off across accuracy/error metrics, latency, and energy efficiency, confirming that the proposed framework benefits from the interaction of its edge–cloud split and collaborative federated optimization.

The radar charts in Fig. 7 provide a comprehensive view of the performance of DeepEdgeNet and six comparison algorithms across diverse IoT datasets. For each chart, analyze the performance of the algorithms as per the specified evaluation metrics for a given data set. DeepEdgeNet is found to outperform almost all comparison algorithms over the evaluated metrics with significant but not limited focus on high accuracy and efficient energy consumption, coupled with low-latency metrics. IoT-based environmental modeling has attracted a lot of research interest in recent years because of the efficient usage of resources available at the edge of IoT systems and because of the possibility of gaining real-time insights to complex phenomena and the behavior of the environment. Modeling is the backbone of IoT-based environmental modeling. Environmental models should provide the required information on time while utilizing the resources available from the sensor networks to the fullest extent possible. Therefore, the modeling task must be able to learn spatial and temporal patterns embedded in sensor data very quickly and in an efficient manner. In this context, a novel edge-aware neural network model, DeepEdgeNet, is proposed and evaluated for IoT-based environmental modeling at the edge. Experiments are conducted with four different sensor datasets (Air Quality and Individual Household Power and US Drought for regression and EuroSAT

for classification). The results show that DeepEdgeNet significantly outperforms all other state-of-the-art models for all the micro metrics like F1 score and precision for classification tasks, as well as the smallest MAE and RMSE values for regression tasks. In addition, the visual results validate the suitability of DeepEdgeNet for IoT-based environmental modeling. Fig. 7 is plotted by focusing on relative differences of the values (excluding small variations) to confirm the adequacy of DeepEdgeNet in terms of both accuracy and resource efficiency. Fig. 8 and 9 confirm the winning position of DeepEdgeNet against all the other five baseline algorithms in both the Air Quality and EuroSAT datasets, achieving the highest average scores for all the predictive metrics, including accuracy, precision, recall, and F1 score. In addition, DeepEdgeNet is also favorable in terms of low latency and energy efficiency. Therefore, DeepEdgeNet is preferred for the Air Quality data set because of its fast inference and lower resource requirements, which is highly preferred in real-time pollution alert systems, while DeepEdgeNet is preferred for the EuroSAT data set because of its capacity to handle the high-dimensional remote sensing data efficiently.

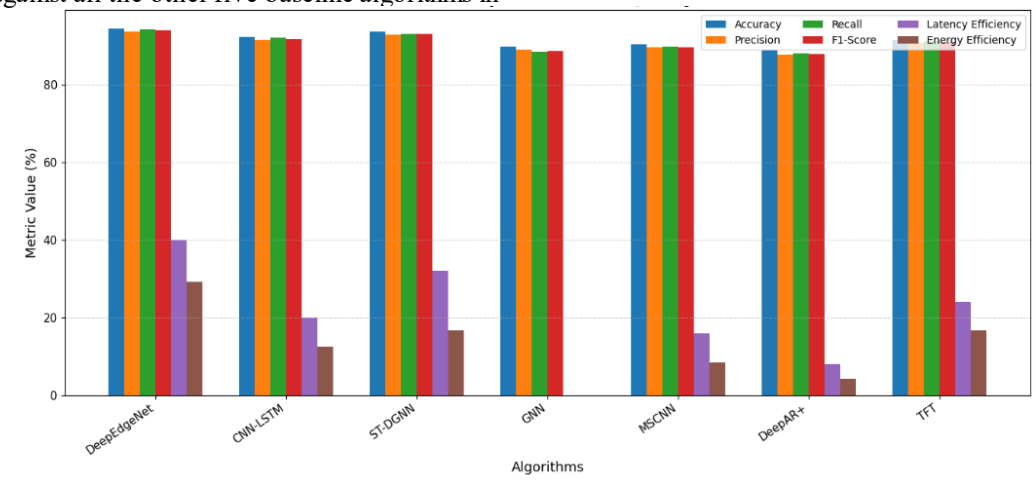


Fig. 8. Performance comparison of DeepEdgeNet and six baseline algorithms on the Air Quality dataset.

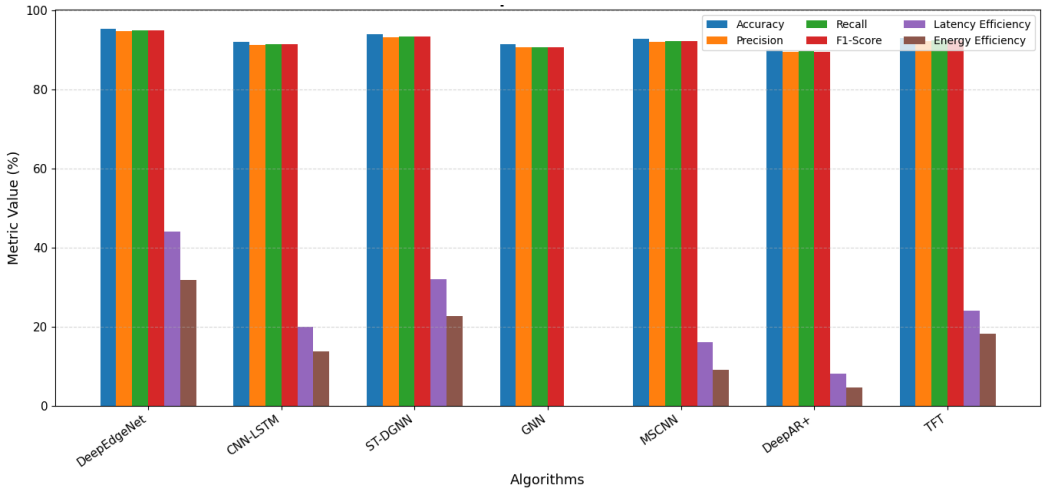


Fig. 9. Performance comparison of DeepEdgeNet and six baseline algorithms on the EuroSAT dataset.

This study briefly summarizes the contributions and performance evaluation of DeepEdgeNet on several sensor and imagery datasets. Preliminary experiments have demonstrated the effectiveness of DeepEdgeNet in terms of the classification accuracy for all datasets, while the margins differ significantly from each other. Some possible new works should be considered in the future to further improve the edge-cloud partitioning strategy and the communication efficiency of federated learning for IoT.

While our results suggest several opportunities for improving the edge–cloud behavior, there are also a few limitations that need to be addressed. Currently, all our experiments are based on simulation. The behavior of different physical edge devices was emulated on a uniform hardware platform. Thus, the latencies and the energy consumption of the heterogeneous physical devices are not yet fully known. Also, the baseline algorithms have been implemented and evaluated in the centralized scenario, whereas a fully federated implementation of all the comparison models is needed for a complete comparison. Another limitation is that our experiments are based on four different environmental monitoring datasets, and that it will be necessary to validate the robustness of our studies with respect to possible disturbances (such as noise and communication faults) as well as hardware heterogeneity in

large-scale IoT systems, which have not yet been investigated. Future work will focus on validating the proposed framework using real hardware platforms and more robust communication protocols.

VII. Conclusion

DeepEdgeNet: An Edge-Fed Deep Learning Framework toward Smart IoT for Environmental Monitoring. DeepEdgeNet is an edge-fed deep learning framework for enabling efficient, scalable, and secure IoT-based environmental monitoring systems. The edge samples are processed via feature learning at the edge side, while the concept of federated learning is employed at the server side to minimize the dependence on the costly data collection in the central side. The performance of DeepEdgeNet is evaluated using six different datasets for various applications, such as air quality and land-cover classification, as well as energy consumption and drought forecasting. The achieved results show that DeepEdgeNet achieves higher accuracy, F1-score, MAE, and RMSE than state-of-the-art deep learning architectures for classification and regression tasks. In addition, the proposed framework provides lower latency and energy consumption required for IoT-based environmental monitoring systems. Although the current implementation of DeepEdgeNet is performed in the simulated edge–cloud environment, the performance trends obtained in this work confirm the applicability of DeepEdgeNet for real-time environmental monitoring that needs immediate analysis along with efficient resource utilization. In future work, we plan to evaluate DeepEdgeNet on real hardware-based heterogeneous edges. In addition to Google LeNet, we plan to add other baseline models to federated learning. We also plan to implement an adaptive communication mechanism to minimize the overheads.

References

[1] B. Katie, “Internet of Things (IoT) for Environmental Monitoring,” 2024. [Online]. Available: www.carijournals.org

[2] X. Sun and N. Ansari, “EdgeIoT: Mobile Edge Computing for the Internet of Things,” IEEE Communications Magazine, vol. 54, no. 12, pp. 22–29, Dec. 2016, doi: 10.1109/MCOM.2016.1600492CM.

[3] S. Chen et al., “Internet of Things Based Smart Grids Supported by Intelligent Edge Computing,” IEEE Access, vol. 7, pp. 74089–74102, 2019, doi: 10.1109/ACCESS.2019.2920488.

[4] B. Omoniwa, R. Hussain, M. A. Javed, S. H. Bouk, and S. A. Malik, “Fog/edge computing-based IoT (FECIoT): Architecture, applications, and research issues,” IEEE Internet Things J, vol. 6, no. 3, pp. 4118–4149, Jun. 2019, doi: 10.1109/JIOT.2018.2875544.

[5] B. Chen, J. Wan, A. Celesti, D. Li, H. Abbas, and Q. Zhang, “Edge Computing in IoT-Based Manufacturing,” IEEE Communications Magazine, vol. 56, no. 9, pp. 103–109, 2018, doi: 10.1109/MCOM.2018.1701231.

[6] T. Okot, P. Madrigal-Mendez, and D. Solorzano-Arias, “The Internet of Things (IoT) for Sustainability: A Framework for Costa Rica,” Journal of Technology Management & Innovation, vol. 18, no. 4, 2023, [Online]. Available: http://jotmi.org

[7] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, “Intelligent Edge Computing for IoT-Based Energy Management in Smart Cities,” IEEE Netw, vol. 33, no. 2, pp. 111–117, Mar. 2019, doi: 10.1109/MNET.2019.1800254.

[8] J. Li et al., “A Secured Framework for SDN-Based Edge Computing in IoT-Enabled Healthcare System,” IEEE Access, vol. 8, pp. 135479–135490, 2020, doi: 10.1109/ACCESS.2020.3011503.

[9] M. Ramadane, S. Meyer, and D. Bohnet, “Environmental Impact Assessment of IoT Devices: A Graph-based Approach,” in Procedia Computer Science, Elsevier B.V., 2024, pp. 338–347. doi: 10.1016/j.procs.2024.05.039.

[10] A. Salam, “Internet of things for environmental sustainability and climate change,” in Internet of Things, Springer, 2020, pp. 33–69. doi: 10.1007/978-3-030-35291-2_2.

[11] R. Verma, R. Raina, V. Garg, S. Gautam, and S. Kumar, “An IoT-Based Biodiversity Monitoring System,” in 2024 IEEE International Conference on Omni-Layer Intelligent Systems, COINS 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/COINS61597.2024.10622521.

[12] T. L. Narayana et al., “Advances in real time smart monitoring of environmental parameters using IoT and sensors,” Heliyon, vol. 10, no. 7, p. e28195, Apr. 2024, doi: 10.1016/j.heliyon.2024.e28195.

[13] S. Piramuthu, “IoT, Environmental Sustainability, Agricultural Supply Chains,” in Procedia Computer Science, Elsevier B.V., 2022, pp. 811–816. doi: 10.1016/j.procs.2022.08.098.

[14] M. L. Nirere, K. Jayavel, and A. Ngenzi, “IoT-BASED CLIMATE CHANGE PREDICTION SYSTEM,” in ACM International Conference Proceeding Series, Association for Computing Machinery, Feb. 2023, pp. 227–233. doi: 10.1145/3587828.3587862.

[15] C. Campolo, G. Genovese, G. Singh, and A. Molinaro, “Scalable and interoperable edge-based federated learning in IoT contexts,” Computer Networks, vol. 223, p. 109576, Mar. 2023, doi: 10.1016/j.comnet.2023.109576.

[16] Y.-L. Hsu, C.-F. Liu, H.-Y. Wei, and M. Bennis, “Optimized Data Sampling and Energy Consumption in IIoT: A Federated Learning Approach,” IEEE Transactions on Communications, vol. 70, no. 12, pp. 7915–7931, Dec. 2022, doi: 10.1109/TCOMM.2022.3216353.

[17] Qamar H. Naith, H. Mancy, An IoT-Enabled Hybrid DRL-XAI Framework for Transparent Urban Water Management,CMES - Computer Modeling in Engineering and Sciences,Volume 144, Issue 1,2025,pp. 387-405, doi:10.32604/cmes.2025.066917.

[18] N. Abd El-Mawla, M. Badawy, and H. Arafat, “IoT for the Failure of Climate-Change Mitigation and Adaptation and IIoT as a Future Solution,” World Journal of Environmental Engineering, vol. 6, no. 1, pp. 7–16, Apr. 2019, doi: 10.12691/wjee-6-1-2.

[19] B. Lim, S. Arık, N. Loeff, and T. Pfister, “Temporal Fusion Transformers for interpretable multi-horizon time series forecasting,” Int J Forecast, vol. 37, no. 4, pp. 1748–1764, Oct. 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[20] Z. Wu, F. Chen, G. Long, S. Pan, C. Zhang, and P. S. Yu, “A Comprehensive Survey on Graph Neural Networks for Knowledge Graphs,” IEEE Access, vol. 10, pp. 75729–75741, 2022, doi: 10.1109/ACCESS.2022.3191784.

[21] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in International Conference on Artificial Intelligence and Statistics, 2016. doi: 10.1109/JIOT.2018.2805263.

[22] H. Alhichri, N. Alajlan, Y. Bazi, and T. Rabczuk, “Multi-Scale Convolutional Neural Network for Remote Sensing Scene Classification,” in IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 2018, pp. 1–5.

[23] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, “DeepAR: Probabilistic forecasting with autoregressive recurrent networks,” Int J Forecast, vol. 36, no. 3, pp. 1181–1191, Jul. 2020, doi: 10.1016/j.ijforecast.2019.07.001.

[24] Bing Yu, Haoteng Yin, and Zhanxing Zhu, “Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting,” in International Joint Conferences on Artificial Intelligence Organization (IJCAI-18), Jul. 2018, pp. 3634--3640. doi: 10.24963/ijcai.2018/505.

[25] D. Martin Ward Powers, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION,” Article in Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011, doi: 10.9735/2229-3981.

[26] G. Premsankar, M. Di Francesco, and T. Taleb, "Edge Computing for the Internet of Things: A Case Study," IEEE Internet Things J, vol. 5, no. 2, pp. 1275–1284, Apr. 2018, doi: 10.1109/JIOT.2018.2805263.

[27] S. Mittal, "A survey of techniques for improving energy efficiency in embedded computing systems," International Journal of Computer Aided Engineering and Technology, vol. 6, no. 4, pp. 440–459, 2014, doi: 10.1504/IJCAET.2014.065419ï.

[28] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–487, 2013, doi: 10.1561/0400000042.

[29] S. Vito, "Air Quality," UCI Machine Learning Repository.

[30] G. Hebrail and A. Berard, "Individual Household Electric Power Consumption," UCI Machine Learning Repository.

[31] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 12, no. 7, pp. 2217–2226, 2019, doi: https://www.kaggle.com/code/nilesh789/land-cover-classification-with-eurosat-dataset.

[32] Centre for Research on the Epidemiology of Disasters (CRED), "Emergency Events Database (EM-DAT)," Université catholique de Louvain, Belgium, 2024.

[33] ADITYA KADIWAL, "Water Potability Dataset," https://www.kaggle.com/datasets/adityakadiwal/water-potability.

[34] NASA POWER Project and authors of the US Drought Monitor, "Predict Droughts using Weather & Soil Data," Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy., 2023, doi: https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data.

Machine Learning Evaluation of the HiTar-2024 Dataset for Intrusion Detection in Smart Manufacturing Environments

Adeeb Alhomoud
College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

***Abstract*—This study presents an investigation of the HiTar-2024 dataset performed in terms of the distribution of label attack types and the distribution of attacks by protocol, normal, and Denial of Service (DoS) connections over time. The investigation carried out a performance evaluation of the HiTar-2024 dataset using a machine learning approach to classify benign and malicious activities, based on BayesNet, Logistic, IBk, Multiclass, PART, and J48 classifiers. It was found that the HiTar-2024 dataset can serve as a training set for an anomaly-based intrusion detection system (IDS) in a smart manufacturing environment to detect normal and malicious activities. Furthermore, an anomaly-based IDS using the HiTar-2024 dataset is able to group malicious activities into Probing, Remote-to-Local, User-to-Root, and DoS attacks.**

Keywords*—Supervised learning; intrusion detection system; security; smart manufacturing

I. Introduction

Today, intrusion detection systems (IDS) play a crucial role in detecting and stopping sophisticated attacks and malware propagation, especially in the Internet-of-Things (IoT) environment known as resource constrained devices. The smart grid is a particular case of an IoT environment in which any activity disruption leads to catastrophic damage to the fabrication process of smart manufacturing, market, service provider, and energy distribution systems [1]. The IDS taxonomy in the IoT environment contains four categories: signature-based [2], anomaly-based [3], specification-based [4], and hybrid detection mechanisms [5]. An anomaly-based IDS relies on machine learning (ML) and deep learning (DL) techniques. An optimized IDS ML based on a support vector machine (SVM) has been proposed by fitting the SVM kernel function and the complexity parameter, leading to a maximal value of the accuracy of the model [6]. A DL classifier called a Bidirectional long short-term memory (Bi-LSTM) model has achieved accuracy of up to 97.93% [7].

The taxonomy of attacks based on KDD'99 is composed of four classes: Probing, Remote-to-Local (R2L), User-to-Root (U2R), and Denial-of-Service (DoS) attacks [8]. In a Probing attack, the hacker will gather the maximum information possible about the target IoT victim vulnerability [9]. The hacker will then exploit this vulnerability to gain R2L access to the target IoT device or network gateway [10]. In addition, the aim of a U2R attack is, for instance, to gain root privileges through a memory overflow mechanism [11]. Finally, a hacker can stop or degrade the performance of a target IoT network through a DoS or Distributed DoS (DDoS) attack [12].

In Industrial IoT (IIoT) environments, the creation and thorough analysis of dedicated datasets is fundamental to developing an effective IDS. Standard network-security datasets, such as those developed for traditional IT networks, often fail to capture the unique characteristics of IIoT architectures, which include constrained devices, heterogeneous sensors, specialized communication protocols, and cyber-physical interactions. Recent research by Essop et al. [13] emphasizes that many existing IoT and IIoT datasets fail to capture the diversity of real-world conditions, as they often lack events that encompass multiple benign and attack scenarios. Furthermore, these datasets typically omit critical contextual information, such as sensor measurement data, network-level traffic characteristics, and device behavior within operational environments. The introduction of the Canadian Institute for Cybersecurity Advanced Persistent Threat for IIoT (CICAPT-IIoT) dataset addresses this gap by including network logs and provenance data across more than 20 advanced persistent threat techniques in an IIoT testbed [14]. Through dataset construction and detailed analysis of attack distributions, IoT device behavior, protocol usage distribution, and temporal patterns, researchers can better understand threat vectors related to the IIoT environment. This process could enable the design of IDS models based on ML that are both accurate and practical in various applications, such as smart manufacturing, energy distribution, and critical infrastructure settings. Without the existence of those targeted datasets, the IDS models generated will suffer from poor generalizability, class imbalance, or misalignment with real-world IIoT networks. The main contributions of this study are summarized as follows:

- An investigation of the HiTar-2024 dataset generated from a smart manufacturing simulation environment.
- A HiTar-2024 dataset performance evaluation based on BayesNet, Logistic, Instance-Based k-Nearest Neighbors (IBk), Multiclass, partial decision trees (PART), and J48 classifiers.
- The HiTar-2024 dataset can serve as a training set to detect Probing, R2L, U2R, and DoS attacks before its deployment online in a smart manufacturing environment.

The rest of the study is composed into four sections: Section II presents the state-of-the-art of the existing IIoT dataset in the literature. Section III describes the HiTar-2024 dataset by considering its most important features. Section IV investigates a performance evaluation of the HiTar-2024 dataset and presents a confusion matrix of the classifiers used. Section V presents the results and discussion. Section VI draws the conclusion.

II. State-of-the-Art

Table I compares IIoT datasets in the literature as a function of year creation, dataset name, number of features, attack classes, and a brief description.

TABLE I. IIoT Datasets in the Literature

Year	Dataset Name	Number of Features	Attack Classes	Description
2020	TON_IoT [15]	127	Normal, DDoS, DoS, Injection, MiM, Password, XSS, Scanning.	A telemetry dataset is proposed for both IoT and IIoT applications.
2021	WUSTL-IIOT-2021 [16]	41	Command Injection, DoS, Reconnaissance, Backdoor.	WWUSTL-IIoT-2021 consists of network data of Industrial Internet of Things to be used in cybersecurity research.
2022	Edge IIoTset [17]	61	DoS/DDoS attacks, Information gathering, MiM attacks, Injection attacks, Malware attacks.	The Edge-IIoTset dataset generated could be used by an IDS based on centralized and federated learning.
2024	CICAPT-IIoT [18]	140	Collection, Exfiltration, Command and Control, Persistence, Discovery, Credential Access, Lateral Movement and Defence Evasion.	An APT attack dataset captured within the IIoT environment.
2024	HiTar-2024 [19]	39	Probing, R2L, U2R and DoS	HiTar-2024 is a dataset generated from the AREZZO environment that represents a simulated industrial manufacturing process
2025	DataSense: CIC IIoT [20]	94	Benign, DoS, DDoS, MiTM, Mirai, Bruteforce	A realistic IIoT testbed dataset with synchronized sensor and network data for anomaly detection.

Important IIoT datasets that are currently available in the literature are compared in this work along with the main distinctions between them, such as the number of features extracted, the variety of attacks that are possible, and their suitability for cybersecurity research in the smart manufacturing field. A summary of the early attempts to create comprehensive datasets is provided. One of the first comprehensive datasets was the TON_IoT [15], which had 127 features and covered a wide range of attack types, such as DDoS, DoS, injection, and password attacks, making it suitable for anomaly detection based on telemetry for IoT/IIoT networks.

The 41 features of WUSTL-IIoT-2021 [16], on the other hand, cover injection, DoS, reconnaissance, and backdoor attacks, all of which are essential for assessing targeted IIoT attacks. The Edge IIoTset dataset contains 61 new features and could be used by an IDS based on centralized and federated learning [17]. The CICAPT-IIoT dataset was introduced in 2024 [18]. The latter dataset contains 20 kinds of attacks that can be carried out in an IIoT environment, including collection, exfiltration, command and control, persistence, discovery, credential access, lateral movement, and defence evasion classes. Another, HiTar-2024, is a dataset generated from AREZZO environment LOG files that represents a simulated industrial manufacturing process in an IIoT context. The HiTar-2024 dataset includes 39 features, 15,842 instances, and four attack classes: Probing, R2L, U2R, and DoS attacks [19]. Furthermore, the HiTar-2024 was created following a rigorous approach involving data preprocessing, feature extraction, the class imbalance with SMOTE, and a test option that was employed for model robustness. Finally, the DataSense CIC-IIoT Dataset 2025, also developed by the CIC, represents one of the most recent and comprehensive datasets designed for intrusion detection in IIoT environments. Built from a realistic IIoT testbed, it integrates both sensor telemetry and network traffic data, providing a synchronized view of industrial device behavior and communication flows [20].

III. Description of the Dataset

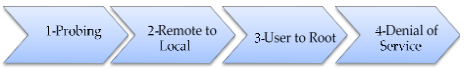


Fig. 1. Attack class sequence order

Fig.1 represents the chronological attack class order used in HiTar-2024 generation. The process started by collecting information about the target victim (probing class). In the attack scenario, NMAP commands were used to perform a probing attack. Subsequently, a remote connection to the target AREZZO system was attempted (R2L class). In addition, Rlogin, Secure Shell (SSH), and Telnet commands were examined as the R2L attack class. A U2R class is a tentative attempt to gain Root permission. Sudo, Su, and FTP commands were then investigated to generate U2R attacks. Finally, a DoS attack was launched in order to stop or degrade a service or network activity related to the AREZZO environment (DoS class). The DoS attack was investigated through the Hping3 command.

The following section presents the distribution of label attack types and the distribution of attacks by protocol, normal, and DoS connections over time in the HiTar2024 dataset using the JupyterLab environment. The pie chart in Fig. 2 shows the distribution of label attack types in the HiTar2024 dataset. Each

slice represents how frequently a specific behavior or attack type occurs in the dataset. As shown above, Probing attacks represent 65.6% of the total attacks, Normal behavior represents 25.6% of all traffic, R2L attacks are 4.9% of the total traffic, U2R attacks are 3.1% of the total attacks, and DoS attacks represent 0.8% of the global collected traffic. This indicates that the majority of the records in the dataset involve scanning or probing activities, whereby an attacker tries to gather information about the system (e.g., port scanning or network mapping). The dataset is highly imbalanced, with one class (Probing_Attack) dominating. As a consequence, a more specific technique (e.g., class weighting, oversampling rare attacks such as a DoS, or undersampling probing attacks) is necessary to avoid biased detection models.

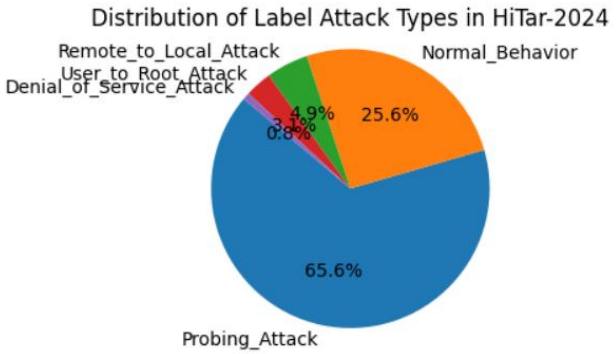


Fig. 2. Distribution of label attack types in the HiTar-2024 dataset.

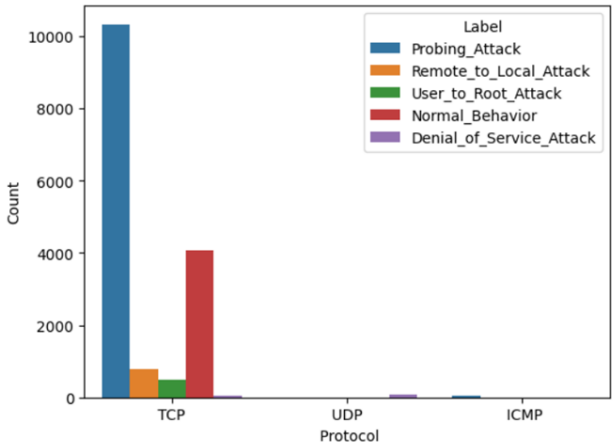


Fig. 3. Distribution of attacks by protocol in the HiTar-2024 dataset.

Fig. 3 shows the distribution of different attack categories and normal traffic across network protocols (Transmission Control Protocol [TCP], User Datagram Protocol [UDP], and Internet Control Message Protocol [ICMP]) within the HiTar-2024 dataset. The results show a strong imbalanced usage of protocol types, indicating that the majority of both benign and malicious activities are transmitted over TCP. Specifically, Probing attacks are the most frequent, exceeding 10,000 TCP instances, suggesting extensive scanning or probing activity targeting IoT devices. R2L and U2R attacks also occur mainly over TCP, although at considerably lower frequencies, which aligns with these attacks typically requiring a stable connection to be established in order to exploit system vulnerabilities and weaknesses. Finally, Normal traffic is mainly transmitted over TCP, with 4000 instances, reinforcing the central role of TCP in typical IIoT communication patterns. To draw a conclusion, the imbalance across protocols indicates that IIoT intrusion detection models should incorporate knowledge of protocol features, particularly identifying TCP traffic patterns.

Fig. 4 shows the Normal and DoS connections over time in the HiTar-2024 dataset. The pic connections represent the DoS attacks when compared to Normal activity (without pic connections).

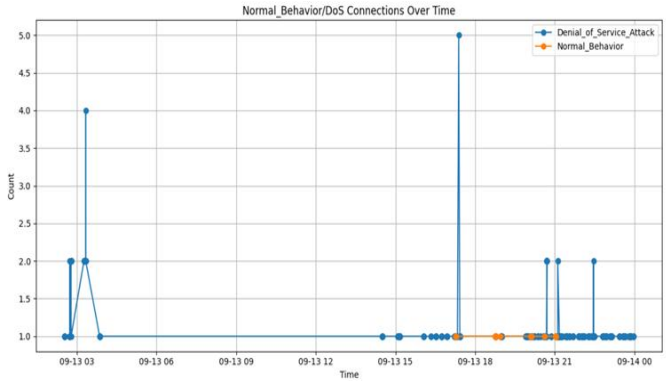


Fig. 4. Normal and DoS connections over time in the HiTar-2024 dataset.

As seen in Fig. 4, the various spikes represent DoS attacks, and the line below represents Normal activity. Fig. 4 indicates a high-intensity attack phase, which is likely to correspond to a targeted flooding attempt. In contrast, Normal behavior appears only sporadically during this interval. Between 06:00 and 15:00 on 09-13, the traffic stabilizes, and DoS activity is minimal, suggesting either a paused attack or normal operational conditions in the IIoT network. However, around 18:00, another sharp spike in DoS activity is observed, indicating a second attack wave, again with significantly higher frequency than the normal traffic. Approaching the later period (around 20:00 to 23:00), both Normal and DoS connections appear intermixed, but DoS packets still dominate in terms of count. This behavior may indicate either background scanning/flooding attempts or a sustained low-rate persistent DoS attack intended to evade detection.

IV. Dataset Performance Evaluation

A structured comparison of the different classifiers tested for the HiTar-2024 dataset is shown in Table II. BayesNet, Logistic regression, IBk, Multiclass classification techniques, PART, and J48 (also known as the C4.5 algorithm) are included in the table. The computational complexity, main benefits in the classification context, and potential drawbacks that could affect the suitability of each classifier for IIoT threat detection are described in the table.

TABLE II. CLASSIFIERS USED IN HITAR-2024 PERFORMANCE EVALUATION

Classifier Name	Description	Complexity	Advantages	Disadvantages
BayesNet [19,22,24]	A probabilistic graphical model that uses Bayesian networks for classification.	Low	- Efficient for small datasets - Handles missing data well - Interpretable results	- Assumes conditional independence - Less effective with complex, high-dimensional data
Logistic [19,21,23,24]	A statistical model that predicts categorical outcomes using a logistic function.	Low to Medium	- Works well with linear decision boundaries - Robust against overfitting when regularization is applied	- Struggles with non-linearly separable data - Performance drops with highly imbalanced datasets
IBk [19,24]	A lazy learning algorithm that classifies based on the majority class of nearest neighbors.	High	- Simple and intuitive - Effective for non-linear relationships - No training phase, adapts dynamically	- Computationally expensive for large datasets - Sensitive to noise and irrelevant features
Multiclass [19,24]	A classification method that extends binary classifiers to handle multiple classes simultaneously.	Medium to High	- Efficient for multi-label problems - Can be integrated with various base classifiers	- Performance depends on the decomposition method chosen - Computational cost increases with class complexity
PART [19,21,24]	A rule-based classifier that builds decision rules from partial decision trees.	Medium	- Generates interpretable rules - Less prone to overfitting compared to full decision trees	- Can be computationally expensive - Rule generation may not always be optimal
J48 [19,21,24]	A decision tree-based classifier that recursively splits data based on attribute values.	Medium to High	- Easily interpretable - Handles both numerical and categorical data - Can handle missing values	- Prone to overfitting - Can become complex and hard to optimize

The following classifier configurations were used to perform system performance evaluation:

weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.

weka.classifiers.functions.Logistic ---R 1.0E-8 -M -1 -num-decimal-places 4

weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

weka.classifiers.rules.PART -C 0.25 -M 2

weka.classifiers.meta.MultiClassClassifier -M 0 -R 2.0 -S 1 -W

weka.classifiers.trees.J48 -C 0.25 -M 2

The methodology combined three effective techniques to address overfitting in the tabular data in the HiTar-2024 dataset, including balancing the minority class using SMOTE (Synthetic Minority Over-sampling Technique), feature selection by selecting an attribute evaluator, and a search method and k-fold cross-validation in test option processing. Regarding feature selection, the CfsSubsetEval and BestFirst were selected as the attribute evaluator and search method, respectively. The choice of the CfsSub-setEval feature selection method was motivated by this suitability for tabular data, it produces a compact, nonredundant subset, and improves model generalization.

Furthermore, imbalanced classes are handled by SMOTE, such as the U2R and DoS attack classes. The following setting is used in the preprocessing supervised filter based on the instance:

weka.filters.supervised.instance SMOTE -C 0 -K 5 -P 400.0 -S 1

Since the minority class DoS has a number of instances of 130, the percentage parameter of SMOTE is selected to 400% to reach the number of U2R instances equal to 491.

A k-fold cross-validation test option was also applied, partitioning the data into k subsets (or folds), training the model on k-1 subsets, and testing it on the held-out subset. The simulation was carried out for k equal to 10 for the cross-validation test option.

weka.attributeSelection.CfsSubsetEval -P 1 -E 1

CfsSubsetEva evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

weka.gui.attributeSelection.BestFirst -D 1 -N 5

BestFirst searches the space of attribute subsets by using greedy hill climbing augmented with a backtracking facility.

A comprehensive set of experiments with common ML models using BayesNet, Logistic, IBk, Multiclass, PART, and J48 with the HiTar-2024 dataset demonstrates high accuracy, precision, recall, and F1-scores, as well as around 99% accuracy [19]. For example, the authors in [19] made a comparison between Simple Logistic, PART, and J48 classifiers using the NSL-KDD dataset. They also examined the detection of different attack types involving DoS, R2L, U2R, and Probing attacks and considered time when building the model. The authors in [22] tested the BayesNet classifier using NSL-KDD and found higher true positive (TP) and lower false positive (FP) rates when compared to Naïve Bayes and updateable versions. Other authors proposed Logistic regression, decision trees, k-Nearest Neighbors, and Naïve Bayes classifiers on the WUSTL-IIOT-2021 dataset [23]. The latter study gives a context for how IBk and Logistic classifiers perform under IIoT network traffic data conditions.

Table III presents a performance comparison of classifiers for HiTar-2024 as a function of accuracy, Kappa statistic, mean absolute error (MAE), root mean square error (RMSE), and relative absolute error (RAE) metrics.

As shown in Table III, all classifiers demonstrate remarkably high accuracy, exceeding 98%, which indicates that the proposed dataset and feature selection approach enable highly reliable classification. Among the models evaluated, the IBk algorithm achieved the highest accuracy, reaching 99.993%, accompanied by a Kappa statistic of 0.999, MAE of 0.0001, and RMSE of 0.005. These results highlight the robustness and consistency of the IBk model, confirming its superior ability to minimize both MAE and RMSE errors. The Logistic regression model also achieved competitive results, with an accuracy of 99.974%, a Kappa of 0.9995, MAE of 0.0001, and RMSE of 0.0101. The J48 decision tree and Multiclass classifiers followed closely, each exceeding 99.9% accuracy, and demonstrating excellent agreement levels with a Kappa of around 0.999.

TABLE III. CLASSIFIER PERFORMANCE COMPARISON FOR HITAR-2024

Classifier	Accuracy (%)	Kappa Statistic	Mean Absolute Error	Root Mean Square Error	Relative Absolute Error
BayesNet	98.668	0.973	0.005	0.069	2.471
Logistic	99.974	0.9995	0.0001	0.0101	0.053
IBk	99.993	0.999	0.0001	0.005	0.0623
PART	99.899	0.998	0.019	0.2903	6.2892
Multiclass	99.949	0.999	0.0004	0.0143	0.2094
J48	99.968	0.999	0.0002	0.011	0.123

In contrast, although the PART and BayesNet classifiers also showed good results with an accuracy of 99.899% and 98.668%, respectively, they showed higher error values regarding an RMSE equal to 0.2903 and 0.069 and an RAE of 6.2892 and 2.471, suggesting a slightly lower prediction precision compared to the top-performing IBk model. Nonetheless, their Kappa statistics of above 0.97 confirm consistent and reliable performance over classes. In conclusion, the IBk and Logistic regression classifiers can be identified as the most suitable models for the HiTar-2024 dataset, achieving excellent classification with minimal prediction error. These findings confirm that the adopted approach is effective at enhancing classification accuracy and maintaining high robustness and generalization capability.

Table IV presents a performance comparison between recent IIoT datasets and HiTar-2024 in terms of task, typical accuracy, precision, recall, and F1-score.

TABLE IV. CLASSIFIER PERFORMANCE COMPARISON BETWEEN HITAR-2024 AND OTHER IIOT DATASETS

Classifier	Task	Typical accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CICAPT-IIoT	APT-style detection	~93.8	~93	~95	~94
Edge-IIoTset	Multiclass attack detection	95–99.97	~95–99.3	~95–99.3	~95–99.3
DataSense (CIC IIoT)	Binary & multiclass IIoT	~98–99	~98	~98	~98
HiTar-2024	Multiclass attack detection	98.668–99.993	99.1–99.9	99.0– 99.9	99.0–99.9

As shown in Table IV, HiTar-2024 outperforms other recent IIoT datasets (i.e., CI-CAPT-IIoT, Edge-IIoTset, and DataSense (CIC IIoT)) in terms of typical accuracy, precision, recall, and F1-score, motiving its suitability for smart manufacturing.

Next, the performance evaluation was conducted using a confusion matrix for the various classifiers studied in a WEKA simulation environment using the same setting as in [17].

Fig. 5(a) presents a colored confusion matrix heatmap for the BayesNet classifier, in which the predicted values are shown in the columns and the defined values of the dataset in the rows. The confusion matrix for the BayesNet classifier presents the following insights:

- Probing attack: The classifier correctly predicted 10,261 instances of Probing attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- R2L attack: The classifier correctly predicts 754 instances of R2L attacks (representing a TP rate of 87.98%). In addition, 102 instances were predicted as probing attacks and one instance predicted as a U2R attack (representing an FP rate of 12.02%).
- U2R attack: The classifier correctly predicted 481 instances of U2L attacks (representing a TP rate of 94.49%), but 16 instances were predicted as probing attacks and 12 instances as R2L attacks (representing an FP rate of 5.51%).
- Normal behavior: The classifier correctly predicted 4058 instances of normal behavior (representing a TP rate of 99.68%). However, 7 instances were predicted as Probing attacks and 6 as R2L attacks (representing an FP rate of 0.32%).
- DoS attack: The classifier successfully predicted 130 instances of DoS attacks (representing a TP rate of 90.27%). Two instances were predicted as Probing

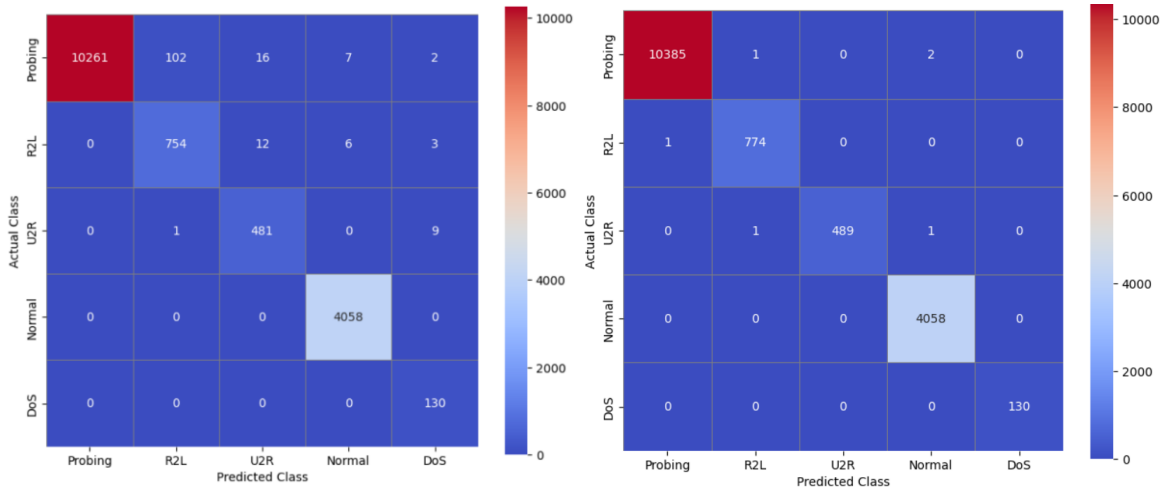
attacks, 3 instances were predicted as R2L attacks, and 9 instances were predicted as U2R attacks (representing an FP rate of 9.73%).

In summary, the BayesNet classifier seems to excel in predicting Probing attacks and Normal behavior, but performs less well in predicting R2L, U2R, and DoS attacks.

Fig. 5(b) presents a colored confusion matrix heatmap for the Logistic classifier, in which the predicted values are shown in the columns and the defined values of the dataset in the rows. The confusion matrix for the Logistic classifier enables the following insights:

- Probing attack: The classifier correctly predicted 10,385 instances of probing attacks (with a TP value equal to 99.99%) and 1 instance was predicted as an R2L attack (with an FP equal to 0.01%).
- R2L attack: The classifier correctly predicted 774 instances of R2L attacks (representing a TP rate of 99.74%). One instance was predicted as a Probing attack and 1 instance as a U2R attack (representing an FP rate of 0.26%).
- U2R attack: The classifier correctly predicted 489 instances of U2R attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- Normal behavior: The classifier correctly predicted 4058 instances of Normal behavior (representing a TP rate of 99.92%). However, 2 instances were predicted as Probing attacks and 1 instance was predicted as a U2R attack (representing an FP rate of 0.08%).
- DoS attack: The classifier correctly predicted 130 instances of DoS attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).

In conclusion, the Logistic classifier seems to excel in predicting U2R and DoS attacks, but performs less well in predicting Probing and R2L attacks and Normal behavior.



(a) (b)

Fig. 5. (a) Colored confusion matrix heatmap for the BayesNet classifier; (b) Colored confusion matrix heatmap for the Logistic classifier.

Fig. 6(a) presents a colored confusion matrix heatmap for the IBk classifier, in which the predicted values are shown in the columns and the defined values of the dataset are shown in the rows. The confusion matrix for the IBk classifier enables the following insights:

- Probing attack: The classifier correctly predicted 10,388 instances of probing attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- R2L attack: The classifier correctly predicted 775 instances of R2L attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- U2R attack: The classifier correctly predicted 491 instances of U2R attacks (with a TP rate equal to 99.79%) and 1 instance was predicted as a DoS attack (with an FP equal to 0.03%).
- Normal behavior: The classifier correctly predicted 4058 instances of Normal activity, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- DoS attack: The classifier correctly predicted 129 instances of U2R attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).

In summary, the IBk classifier seems to excel in predicting Probing, R2L, and DoS attacks and Normal behavior, but performs slightly less well in predicting U2R attacks.

Fig. 6(b) presents the confusion matrix for the Multiclass classifier, in which the predicted values are shown in the columns and the defined values of the dataset in the rows.

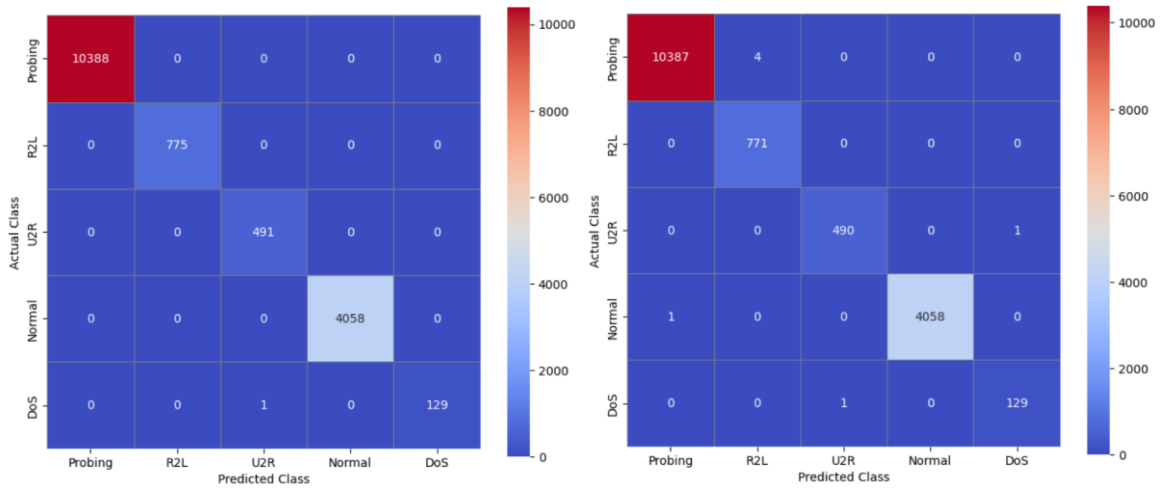


Fig. 6. (a) Colored confusion matrix heatmap for the IBk classifier; (b) Colored confusion matrix heatmap for the Multiclass classifier.

The confusion matrix for the Multiclass classifier presents the following insights:

- Probing attack: The classifier correctly predicted 10,387 instances of Probing attacks (with a TP rate equal to 99.96%) and 4 instances predicted as R2L attacks (with an FP rate equal to 0.04%).
- R2L attack: The classifier correctly predicted 771 instances of R2L attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- U2R attack: The classifier correctly predicted 490 instances of U2R attacks (with a TP rate equal to 99.79%) and 1 instance predicted as a DoS attack (with an FP rate equal to 0.21%).
- Normal behavior: The classifier correctly predicted 4058 instances of U2R attacks (with a TP rate equal to 99.97%) and 1 instance predicted as a Probing attack (with an FP rate equal to 0.03%).
- DoS attack: The classifier correctly predicted 129 instances of U2R attacks (with a TP rate equal to 99.23%) and 1 instance predicted as a U2R attack (with an FP rate equal to 0.73%).

In summary, the Multiclass classifier seems to excel in predicting R2L attacks. However, it performs slightly less well in predicting Probing, U2R, and DoS attacks.

Fig. 7(a) presents a colored confusion matrix heatmap for the PART classifier, in which the predicted values are shown in the columns and the defined values of the dataset in the rows.

The confusion matrix for the PART classifier presents the following insights:

- Probing attack: The classifier correctly predicted 10,387 instances of Probing attacks (with a TP rate equal to 99.96%) and 1 instance predicted as a DoS attack (with an FP rate equal to 0.04%).
- R2L attack: The classifier correctly predicted 772 instances of R2L attacks (with a TP rate equal to 98.84%) and 9 instances predicted as a Probing attack (with an FP rate equal to 1.16%).
- U2R attack: The classifier correctly predicted 491 instances of DoS attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).
- Normal behavior: The classifier correctly predicted 4058 instances of U2R attacks (with a TP rate equal to 99.92%) and 3 instances predicted as R2L attacks (with an FP rate equal to 0.08%).
- DoS attack: The classifier correctly predicted 129 instances of U2R attacks (with a TP rate equal to 98.47%) and 2 instances predicted as Probing attacks (with an FP rate equal to 1.53%).

In summary, based on the confusion matrix above, the PART classifier seems to excel in predicting U2R attacks. However, the number of false alarms is equal to 1, 2, 3, and 9, respectively, for Probing, DoS, Normal, and R2L activities.

Fig. 7(b) shows a colored confusion matrix heatmap for the J48 classifier, in which the predicted values are shown in the columns and the defined values of the dataset in the rows.

The confusion matrix for the J48 classifier presents the following insights:

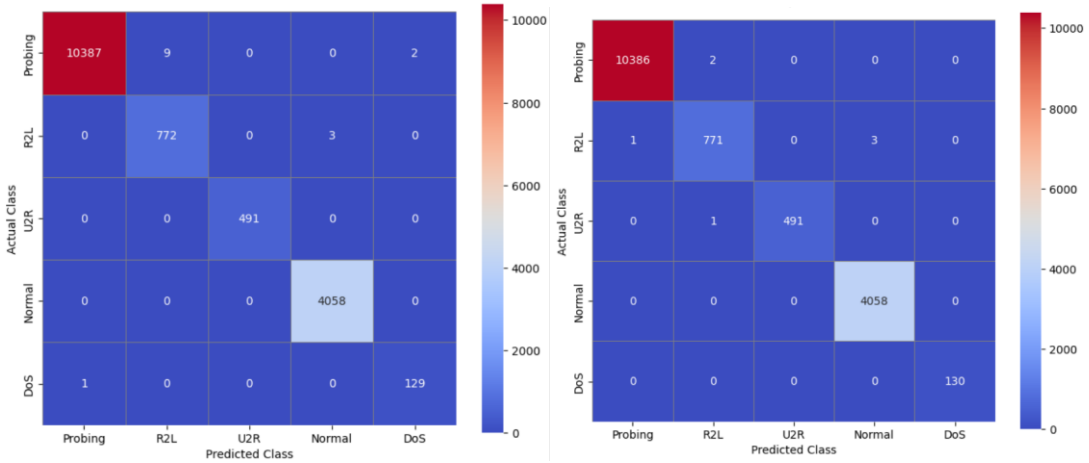
- Probing attack: The classifier correctly predicted 10,386 instances of Probing attacks (with a TP rate equal to 99.99%) and 1 instance predicted as an R2L attack (with an FP rate equal to 0.01%).
- R2L attack: The classifier correctly predicted 771 instances of R2L attacks (with a TP rate equal to 99.74%) and 2 instances predicted as Probing attacks (with an FP rate equal to 0.26%).
- U2R attack: The classifier correctly predicted 491 instances of U2R attacks, with a maximum TP value (i.e.,

equal to 100%) and a minimum FP value (i.e., equal to 0%).

- Normal behavior: The classifier correctly predicted 4058 instances of U2R attacks (with a TP rate equal to 99.92%) and 3 instances predicted as R2L attacks (with an FP rate equal to 0.08%).
- DoS attack: The classifier correctly predicted 130 instances of DoS attacks, with a maximum TP value (i.e., equal to 100%) and a minimum FP value (i.e., equal to 0%).

In summary, the J48 classifier seems to excel in predicting U2R and DoS attacks. However, the number of false alarms is equal to 1, 2, and 3, respectively, for the Probing, R2L, and Normal classes.

Table V compares the various classifiers studied in terms of Probing TP rate, R2L TP rate, U2R TP rate, Normal TP rate, DoS TP rate, average TP rate, and average FP rate.



(a) (b)

Fig. 7. (a) Colored confusion matrix heatmap for the PART classifier; (b) Colored confusion matrix heatmap for the J48 classifier.

TABLE V. TP AND FP RATES PER CLASSIFIERS STUDIED

Classifier	Probing TP Rate	R2L TP Rate	U2R TP Rate	Normal TP Rate	DoS TP Rate	Average TP Rate	Average FP Rate
BayesNet	100	87.98	94.49	99.68	90.27	94.48	5.52
PART	99.96	98.84	100	99.92	98.47	99.43	0.57
Multiclass	99.96	100	99.79	99.97	99.23	99.79	0.21
J48	99.99	99.74	100	99.92	100	99.93	0.07
Logistic	99.99	99.74	100	99.92	100	99.93	0.07
IBk	100	100	99.79	100	100	99.95	0.05

In Table V, the classifier performance is sorted as a function of the average TP rate and FP rate metrics as follows: IBk is the most effective classifier, followed by Logistic regression, then J48 followed by the Multiclass classifier, and the PART classifier is followed by BayesNet. Also as shown in Table V, the best-performing classifier is IBk and the one that performs least well is BayesNet in terms of average TP and FP rates. The FP rate is equal to 0.05 and 5.52 for IBk and BayesNet, respectively. The BayesNet classifier performance is given an overall TP rate of 94.48% and an FP rate of 5.52%.

V. RESULTS AND DISCUSSION

The comparative performance results in Table V above demonstrate clear differences in detection capability and false alarm behavior among the classifiers evaluated in the IIoT intrusion detection context. The BayesNet classifier achieved a reasonable average TP rate of 94.48%, but its relatively high FP rate (5.52%) indicates reduced reliability in high security environments, in which false alerts are costly. The PART classifier showed greater detection performance with an average TP rate of 99.43% and a lower FP rate (0.57%), benefiting from rule-based interpretability, although it remains slightly less accurate than tree-based models. The Multiclass classifier attained a more balanced detection level (99.79% average TP rate) with a low FP rate (0.21%), demonstrating its effectiveness across most attack categories. Among the decision-based methods, both J48 and Logistic regression achieved consistently strong performance (average TP = 99.93%, FP = 0.07%), making them highly suitable for deployment in resource-constrained IIoT edge devices due to their low computational overhead and model transparency. The IBk classifier obtained the highest overall detection accuracy (99.95% average TP rate) and the lowest FP rate (0.05%); however, its high computational cost and memory requirements during inference make it less appropriate for real-time or low-power IIoT nodes, but well-suited for centralized or offline security analytics. Overall, J48 and Logistic emerge as the most practical classifiers for real-time IIoT intrusion detection, whereas IBk is preferable when computational resources are not a constraint. In order to compare

this work with similar research in the literature, the TP rate is shown to be around 95% and the FP rate 4.87% when applying a BayesNet classifier to the NSL-KDD dataset [22]. In resource-constrained IIoT environments, the computational feasibility of ML classifiers depends primarily on inference complexity, memory footprint, and energy consumption. The IBk classifier is generally unsuitable for deployment on IIoT edge devices because it stores the entire training dataset and requires distance computations with all instances during inference, resulting in high memory usage and latency. In contrast, J48 exhibits strong feasibility due to its compact tree representation and low inference complexity, as classification requires only a simple tree traversal. Similarly, PART, a rule-based classifier derived from partial decision trees, offers efficient inference provided that the generated rule set remains limited in size. Multiclass Logistic regression is also viable for edge deployment when trained offline, since inference involves only linear computations over feature vectors, although its cost increases with feature dimensionality and number of classes. BayesNet, although probabilistically expressive, may impose higher memory and computational demands depending on network structure complexity and conditional probability tables. Overall, decision tree-based methods, such as J48, followed by PART and Logistic regression with feature selection, represent the most computationally viable solutions for real-time classification on constrained IIoT edge devices.

Finally, three concrete techniques were implemented to address dataset representativeness and evaluation realism, in terms of data collection (i.e., by oversampling rare attack classes such as the DoS class in HiTar-2024), dataset design (i.e., the HiTar-2024 dataset preserves timestamps in the first attribute to enable realistic splits and analysis), and through better metrics and reporting, such as using a confusion matrix and TP and FP rate per studied class. Although all six classifiers exceed 98% accuracy after SMOTE, the augmentation strategy appears to target only the DoS class, while the rare R2L and U2R classes remain insufficiently represented. Consequently, the reported accuracy may overestimate real-world robustness, and additional cross-dataset or out-of-distribution evaluation is needed to confirm generalization beyond the HiTar-2024 training partition.

VI. Conclusion

In this study, the HiTar-2024 dataset was evaluated using accuracy, Kappa statistic, MAE, RMSE, RAE, and confusion matrix-based performance metrics across several WEKA classifiers, namely BayesNet, Logistic, IBk, Multiclass, PART, and J48. The results indicate that the BayesNet classifier achieved an average TP rate of 94.48% along with an average FP rate of around 5.52%. The PART classifier achieved an average TP rate of 99.43% along with an average FP rate of 0.57%. The Multiclass classifier demonstrated an average TP rate of 99.79% with an average FP rate of 0.21%. An average TP rate of 99.93% along with an average FP rate of 0.07% was shown by the J48 and Logistic classifiers. The IBk classifier recorded the highest value in terms of an average TP rate of 99.95%, along with the lowest value for an average FP rate of 0.05%. These findings demonstrate that the IBk, followed by J48 and Logistic regression, constitute the most suitable set of classifiers for deployment in resource-constrained IIoT environments, offering an effective balance between classification performance and computational efficiency. These models achieve competitive detection accuracy while maintaining low memory consumption and fast inference time, making them particularly well adapted for real-time edge-level implementation.

The presence of high-intensity attack intervals followed by quiet periods in the traffic patterns suggests that the attacker employed intermittent flooding strategies, which is typical behavior in DoS scenarios targeting IIoT networks. Such bursts are designed to disrupt device availability while reducing the likelihood of long-term detection. In addition, the imbalance observed across communication protocols indicates that intrusion detection mechanisms in IIoT environments should incorporate protocol-aware learning features in future work, with particular significance given to TCP traffic behavior being considered as involving the absolute majority of malicious interactions. A potential perspective of this study could be the experiments, which validate detectors in real IIoT environments, the inclusion of advanced attacks such as APTs in the HiTar-2024 dataset, and the use of benchmark hybrid models that combine feature engineered ML, such as CNN/LSTM and CNN/GRU, to strengthen a detection intrusion system in an IIoT environment.

References

[1] Alomari MA, Al-Andoli MN, Ghaleb M, Thabit R, Alkawsi G, Alsayaydeh JAJ, Gaid ASA. Security of Smart Grid: Cybersecurity Issues, Potential Cyberattacks, Major Incidents, and Future Directions. Energies. 2025; 18(1):141. https://doi.org/10.3390/en18010141

[2] Alsoufi, M.A.; Razak, S.; Siraj, M.M.; Nafea, I.; Ghaleb, F.A.; Saeed, F.; Nasser, M. Anomaly-Based Intrusion Detection Systems in IoT Using Deep Learning: A Systematic Literature Review. Appl. Sci. 2021, 11, 8383.

[3] Anzila Saba, Amjad Rehman, Tariq Sadad, Hoshang Kolivand, Saeed Ali Bahaj, Anomaly-based intrusion detection system for IoT networks through deep learning model, Computers and Electrical Engineering, , 2022, Volume 99, 107810, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2022.107810.

[4] S. V. N. Santhosh Kumar and M. Selvi, A. Kannan, A Comprehensive Survey on Machine Learning-Based Intrusion Detection Systems for Secure Communication in Internet of Things, Computational Intelligence and Neuroscience, 2023.

[5] Yazan Otoum, Amiya Nayak, ASIDS: Anomaly and Signature Based IDS for the Internet of Things, Journal of Network and Systems Management, 2021 vol. 29(3),.

[6] A. Alhomoud, An Optimized Network Intrusion Detection System for Attack Detection based on Supervised Machine Learning Models in an Internet-of-Things Environment, International Journal of Advances in Soft Computing & Its Applications, 2023 Vol 15, Issue 2,

[7] Oueslati, N.E., Mrabet, H., Jemai, A. (2025). Intrusion Detection Using an Enhancement Bi-LSTM Recurrent Neural Network Model. In: Ben Hedia, B., Ghazel, M., Monsuez, B. (eds) Verification and Evaluation of Computer and Communication Systems. VECoS 2024. Lecture Notes in Computer Science, vol 15466. Springer, Cham. https://doi.org/10.1007/978-3-031-85356-2_9.

[8] P. Mishra, V. Varadharajan, U. Tupakula and E. S. Pilli, "A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection," in IEEE Communications Surveys & Tutorials, 2019 vol. 21, no. 1, pp. 686-728, Firstquarter, doi: 10.1109/COMST.2018.2847722.

[9] Zahid, M., Bharati, T.S. Enhancing cybersecurity in IoT systems: a hybrid deep learning approach for real-time attack detection. Discov Internet Things 2025 5, 73. https://doi.org/10.1007/s43926-025-00156-y

[10] M. Medwed, V. Nikov, J. Renes, T. Schneider and N. Veshchikov, Cyber Resilience for Self-Monitoring IoT Devices, 2021 IEEE International Conference on Cyber Security and Resilience (CSR), Rhodes, Greece, 2021, pp. 160-167, doi: 10.1109/CSR51186.2021.9527995. keywords: {Virtual machine monitors;Computer architecture;NIST;Hardware;Production facilities;Security;Proposals;Cyber resilience;Internet of Things;Device recovery;Trusted services;Self-monitoring;Attack detection},

[11] Shan, L. (IoT) Network intrusion detection system using optimization algorithms. Sci Rep 2025 15, 21706. https://doi.org/10.1038/s41598-025-04638-5

[12] Anjum, M., Dutta, A.K., Elrashidi, A. et al. GraphFedAI framework for DDoS attack detection in IoT systems using federated learning and graph based artificial intelligence. Sci Rep 2025 15, 28050. https://doi.org/10.1038/s41598-025-10826-0

[13] Essop, I.; Ribeiro, J.C.; Papaioannou, M.; Zachos, G.; Mantas, G.; Rodriguez, J. Generating Datasets for Anomaly-Based Intrusion Detection Systems in IoT and Industrial IoT Networks. Sensors 2021, 21, 1528. https://doi.org/10.3390/s21041528

[14] E. Ghiasvand, S. Ray, S. Iqbal, S. Dadkhah, A. Ghorbani. "Resilience Against APTs: A Provenance-Based IIoT Dataset for Cybersecurity Research".

[15] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood and A. Anwar, TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems, in IEEE Access, 2020 vol. 8, pp. 165130-165150, , doi: 10.1109/ACCESS.2020.3022862.

[16] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain. WUSTL-IIOT-2021 Dataset for IIoT Cybersecurity Research, Washington University in St. Louis, USA, October 2021

[17] Mohamed Amine Ferrag, Othmane Friha, Djallel Hamouda, Leandros Maglaras, Helge Janicke, January "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications: Centralized and Federated Learning, IEEE Dataport, 2022 17, doi: https://dx.doi.org/10.21227/mbc1-1h68.

[18] E. Ghiasvand, S. Ray, S. Iqbal, S. Dadkhah, and A. A. Ghorbani. CICAPT-IIOT: A provenance-based APT attack dataset for IIoT environment, preprint, July 2024.

[19] T. Dhaouadi, H. Mrabet, A. Alhomoud and A. Jemai, An Intrusion Detection System Based on HiTar-2024 Dataset Genera-tion from LOG Files for Smart Industrial Internet-of-Things Environment, Computers, Materials & Continua 2025, Vol. 82, Issue 3,

[20] Firouzi, A.; Dadkhah, S.; Maret, S.A.; Ghorbani, A.A. "DataSense: A Real-Time Sensor-Based Benchmark Dataset for Attack Analysis in IIoT with Multi-Objective Feature Selection." Electronics 2025, 14, 4095

[21] S. Tamy, H. Belhadaoui, N. Rabbah, M. Rifi, Ensemble Learning Based Feature Reduction and Selection Methods For Network Intrusion Detection System, Journal of Theoretical and Applied Information Technology 15th July 2021 - Vol. 99. No. 13 – 2021

[22] Nasir Majeed Mir, Sarfraz Khan, Muheet Ahmed Butt and Majid Zaman, An Experimental Evaluation of Bayesian Classifiers Applied to Intrusion Detection, Indian Journal of Science and Technology, 2016, Volume: 9, Issue: 12, Pages: 1-7

[23] Eid, A.M., Soudan, B., Nassif, A.B. et al. Comparative study of ML models for IIoT intrusion detection: impact of data preprocessing and balancing. Neural Comput & Applic 2024 36, 6955–6972. https://doi.org/10.1007/s00521-024-09439-x

[24] T. Dhaouadi, H. Mrabet and A. Jemai, The HiTar-23 Dataset Construction and Validation For Securing Industrial Internet of Things Environment, 2024 IEEE 27th International Symposium on Real-Time Distributed Computing (ISORC), Tunis, Tunisia, 2024, pp. 1-6, doi: 10.1109/ISORC61049.2024.10551372

ZK-FedMed: Privacy-Preserving Federated Learning for Cardiovascular and Renal Disease Prediction

Haewon Byeon
Department of Future Technology, Korea University of Technology and Education (KOREA TECH), Cheonan, 31253, South Korea

***Abstract*—Protecting patient data confidentiality while enabling collaborative machine learning across distributed healthcare institutions remains a major challenge. This study presents ZK-FedMed, a privacy-preserving federated learning framework that combines CKKS partially homomorphic encryption for gradient protection, Rényi differential privacy with moments-accountant tracking, zk-SNARK-based integrity verification, TabTransformer feature extraction, SCAFFOLD aggregation, and an approximate federated unlearning procedure. The framework was evaluated on two public benchmark datasets: a cardiovascular disease dataset with 69,997 preprocessed records and a chronic kidney disease dataset expanded through SMOTE from 390 unique records to 4,200 balanced records. ZK-FedMed achieved 91.68% precision, 92.28% recall, and 96.73% AUC for cardiovascular prediction and 88.63% accuracy, 88.41% recall, and 94.52% AUC for renal disease classification. Because both datasets are Kaggle-derived and the CKD cohort is heavily augmented, the results are interpreted as benchmark and simulated-federation evidence rather than proof of real-world multi-hospital clinical efficacy. Additional privacy-budget sensitivity analysis showed that stricter budgets such as ε=1.0 and ε=3.0 substantially reduce utility, while ε=10.0 should be understood as a relaxed operational privacy setting rather than a strong practical privacy guarantee. The findings indicate that explicit cryptographic protection and variance-reduced aggregation can improve privacy-aware federated medical prediction while preserving clinically relevant predictive performance under clearly stated data-source and simulation limitations.**

***Keywords*—Federated learning; homomorphic encryption; zero-knowledge proofs; differential privacy; transformer; cardiovascular disease; chronic kidney disease; scaffold; federated unlearning**

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, accounting for approximately 17.9 million deaths annually and representing almost 32% of all deaths worldwide [1]. Chronic kidney disease (CKD) also affects approximately 850 million people worldwide, with an annual economic burden exceeding $48 billion in direct medical costs in the United States alone [2]. Early-stage prediction through machine learning (ML) has shown the potential to reduce adverse clinical outcomes by up to 55% by allowing timely intervention, yet single-institution models suffer from systematic under-representation of diverse patient populations [3]. Multi-institutional collaboration is therefore clinically imperative: analyses across health systems consistently reveal that models trained within a single institution exhibit up to 18% accuracy degradation when deployed in external clinical settings due to demographic bias and equipment heterogeneity [4].

The Health Insurance Portability and Accountability Act (HIPAA) in the United States [5] and the General Data Protection Regulation (GDPR) in the European Union [6] impose strict constraints on centralising patient records, creating a fundamental conflict between the need for large, diverse training sets and the legal obligations governing data sovereignty. Violations attract severe penalties: HIPAA fines range up to $1.9 million per violation category annually, while GDPR penalties can reach €20 million or 4% of global revenue, whichever is greater [7].

Federated learning (FL), first formalized by McMahan et al. [8], offers a compelling paradigm for collaborative model training in which only model updates, rather than raw patient data, are shared with a central aggregation server. However, FL alone does not guarantee privacy: gradient inversion attacks [9] and membership inference [10] have shown that raw gradient transmissions can reconstruct private training samples with high fidelity, rendering unprotected FL models legally and ethically insufficient for healthcare delivery. Furthermore, standard FedAvg [8] assumes independent and identically distributed (IID) data across clients, a principle routinely violated in multi-institutional medical settings where disease prevalence, demographic composition, and clinical protocols differ substantially [11].

Existing defences address these concerns only partially. Differential privacy (DP) [12] introduces calibrated noise to gradient updates at the cost of a 10% to 25% accuracy degradation [13]. Secure multiparty computation (SMPC) [14] provides strong cryptographic guarantees but incurs prohibitive communication overhead (15–25%) and requires synchronous participation. Blockchain-based audit trails offer tamper evidence through hash chaining but do not encrypt the model weights themselves, leaving gradients vulnerable to interception during transmission [15]. Personalization strategies such as K-means clustering [16] mitigate non-IID drift but do not address gradient privacy or Byzantine robustness jointly.

To close these gaps, this study proposes ZK-FedMed (Zero-Knowledge Federated Medicine), a unified framework that integrates five components designed to improve gradient confidentiality, model integrity, privacy accounting, non-IID convergence, and deletion-oriented model updating. The contribution is technical and experimental rather than a claim of direct clinical deployment: the framework is evaluated on public benchmark datasets under simulated institutional partitions.

CKKS homomorphic gradient encryption: Client gradients are encrypted under the partially homomorphic scheme [17] before transmission, enabling the server to perform weighted aggregation directly on ciphertext without ever decrypting individual client updates.

zk-SNARK model integrity verification: After each round of aggregation, a succinct non-interactive argument of knowledge (zk-SNARK) [18] commitment is generated over the global model parameters, allowing any auditor to verify the provenance of the model cryptographically without accessing the weights themselves.

Rényi Differential Privacy with Moments Accountant: Calibrated Gaussian noise is applied to encrypted gradients, and privacy expenditure is tracked using the Rényi DP moments accountant [19]. The revised manuscript explicitly reports the utility loss at stricter ε budgets and clarifies that ε=10.0 is a relatively relaxed operational privacy setting.

TabTransformer Encoder: A multi-head self-attention transformer encoder adapted for tabular EHR data [20] learns nonlinear characteristic interactions among heterogeneous clinical variables, outperforming standard MLP baselines by 1.49–3.21% in precision in both datasets.

SCAFFOLD Aggregation with Federated Unlearning: SCAFFOLD [21] corrects client drift through variance-reduction control variates, improving convergence under non-IID partitions. The integrated federated unlearning module is reframed as an approximate deletion-oriented procedure rather than a complete legal guarantee of GDPR Article 17 compliance.

To the author's knowledge, ZK-FedMed is positioned as a combined privacy-preserving federated learning framework that integrates CKKS homomorphic encryption, zk-SNARK proofs, Rényi DP, TabTransformer, SCAFFOLD aggregation, and approximate federated unlearning in a single healthcare analytics pipeline, while the revised manuscript more explicitly bounds the strength of its privacy, unlearning, and multi-institutional claims.

II. Related Work

A. Federated Learning in Healthcare

McMahan et al. [8] introduced FedAvg, demonstrating that local SGD with periodic model averaging allows effective distributed training. Subsequent work identified serious accuracy degradation under non-IID data: Zhao et al. [11] showed that FedAvg loses 15-20% accuracy when client distributions diverge significantly. FedProx [23] introduced a proximal regularisation term to bound local updates, achieving modest non-IID improvements of 1 to 3%. FedBN [24] addresses the change in feature distribution by keeping batch normalization layers client-local. Rieke et al. [25] provided a comprehensive survey of FL opportunities in healthcare, identifying gradient privacy and heterogeneity of data as the two principal barriers to clinical implementation. Chen et al. [26] developed FedHealth, a federated transfer learning framework for wearable device data that improved accuracy by 1.6% over single-site training but did not address gradient inversion threats. Huang et al. [27] proposed hierarchical FL for multidisease psychiatric diagnosis, achieving a 2.6% improvement over flat federation. SCAFFOLD [21], proposed by Karimireddy et al., corrects client drift using variance reduction control variates, outperforming FedProx by 2–4% on heterogeneous tasks and forming a core component of the ZK-FedMed aggregation strategy (Table I).

B. Privacy-Preserving Machine Learning

Dwork et al. [12] formalised differential privacy (DP), establishing (ε, δ)-DP as the mathematical foundation for privacy-preserving computation. Abadi et al. [28] extended DP to deep learning using the DP-SGD algorithm with a moments accountant, achieving ($\varepsilon = 8$, $\delta = 10^{-5}$)-DP on MNIST with a modest accuracy degradation. Mironov [19] introduced Rényi DP (RDP), which offers strictly tighter privacy composition limits and enables more precise per-round budget accounting per round, a key advantage over classical DP used in ZK-FedMed. Geyer et al. [29] applied DP-SGD to federated learning in healthcare, demonstrating that client-level DP can satisfy regulatory requirements while preserving the utility of the model. Kaissis et al. [30] proposed the PriMIA framework for privacy-preserving medical imaging, reporting only a 1.8% accuracy loss under ($\varepsilon = 8$, $\delta = 10^{-5}$)-DP on chest radiograph classification.

Homomorphic encryption (HE) allows arithmetic operations on encrypted data without decryption, providing strong cryptographic privacy. Gentry [31] introduced the first fully homomorphic encryption scheme; the subsequent CKKS scheme [17] supports approximate floating-point arithmetic, making it practical for neural network gradient aggregation. Mohassel and Zhang [32] demonstrated the feasibility of secure ML inference under HE, although with significant computational overhead. Zhang et al. [33] applied HE-based gradient protection in federated settings, demonstrating 100× overhead compared to plaintext but acceptable latency for periodic model aggregation. The use of CKKS for gradient encryption in ZK-FedMed is directly based on this line of work.

C. Zero-Knowledge Proofs in Distributed Systems

Zero-knowledge proofs (ZKP) [34] allow a prover to demonstrate the truth of a statement to a verifier without revealing any information beyond the validity of the statement. Ben-Sasson et al. [18] introduced zk-SNARKs (succinct noninteractive arguments of knowledge), which produce compact, efficiently verifiable proofs. Groth [35] developed the Groth16 zk-SNARK system that provides the O(1) proof size regardless of circuit complexity, making it suitable for model verification. Recent work has proposed ZKP-based verification for the integrity of ML inference [36], but application to federated learning models remains underexplored. ZK-FedMed extends this line by generating a zk-SNARK commitment over the global model parameters after each aggregation round, enabling any regulatory auditor to verify model authenticity without accessing sensitive weights.

D. Cardiovascular and Renal Disease Prediction

Machine learning approaches for CVD risk stratification have progressed from classical logistic regression achieving 77% accuracy on Framingham cohort data [37] to gradient boosting methods achieving 88-91% AUC on large electronic

health record (EHR) cohorts [38]. Deep learning architectures, including convolutional neural networks applied to ECG waveforms, achieve AUC > 0.97 [39] but require specialised signal acquisition infrastructure. For tabular clinical features, Transformer-based models have recently demonstrated superior performance over tree ensembles by capturing complex nonlinear feature interactions [20]. For the prediction of CKD, Salekin and Stankovic [40] compared multiple classical classifiers on the UCI CKD dataset, with Random Forest achieving 99.1% accuracy under centralised training; however, FL-based CKD prediction with cryptographic privacy guarantees remains largely unstudied. ZK-FedMed addresses this gap by providing the first federated framework that achieves near-centralised performance on both CVD and CKD tasks under formally verifiable privacy constraints.

III. Proposed ZK-FedMed Framework

A. Problem Formulation

Consider N distributed clients (hospitals, clinics, or health networks), each having a private local dataset $D_i = (x, y)$ ni where, $x' \in \mathbb{R}^d$ denotes a d-dimensional clinical characteristic of dimension d and y 0, 1 is the binary diagnostic label. The global objective is to minimise the empirical weighted risk:

$$\theta^* = \arg\min_\theta \sum_{ij1}{}^{P} (n_i/n)\, L_i(\theta; D_i)$$

Subject to: 1) no raw patient data leaves its originating institution; 2) gradient updates are cryptographically protected; 3) the global model satisfies (ε, δ)-RDP; and 4) the trained model carries a verifiable integrity certificate. The challenge is compounded by the following realistic conditions:

Non-IID Distribution: The marginals P(y | x, client i) and the marginals P(x | client i) differ substantially between institutions due to demographics, protocols, and equipment.

Gradient Privacy: Raw gradient transmission enables gradient inversion [9] and membership inference [10]; updates must be encrypted before leaving the client device.

Regulatory Alignment: The framework is designed to support privacy-preserving training and documented deletion-oriented workflows under HIPAA/GDPR motivations, but the unlearning module is treated as approximate empirical unlearning rather than a complete legal certification of GDPR Article 17 compliance.

Auditability: Regulatory bodies require verifiable model provenance without accessing sensitive weight parameters.

B. TabTransformer Encoder Architecture

Unlike the original Transformer [41] designed for sequential data, ZK-FedMed employs a TabTransformer [20] adapted for tabular EHR input. Each categorical characteristic ci 1,..., Ki is embedded in a dense vector $e_i \in \mathbb{R}^{de}$, while continuous features are directly normalised and concatenated. The embedding matrix is $E \in \mathbb{R}^{\{K_i \times d_m\}}$.

The transformer encoder applies L-stacked blocks, each comprising multi-head self-attention (MHSA) and a position-wise feedforward network (FFN):

$$\mathrm{MHSA}(X) = \mathrm{Concat}(\mathrm{head}1, ..., \mathrm{head}h)\mathrm{WOot}$$

$$\mathrm{head}i = \mathrm{Softmax}(Q_i K_i{}^T / \sqrt{d_k})\, V_i$$

where, $Qi = XW_i{}^G$, $K_i = XW_i{}^N$, $V_i = XW_i{}^\beta$ are the query, key and value projections. Each FFN applies two linear transformations with a GELU activation: $\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$. The final representation concatenates transformer-encoded categorical embeddings with normalised continuous features before passing them to a classification head.

C. CKKS Homomorphic Gradient Encryption

After each local training epoch, client i encodes its gradient vector i R under the CKKS scheme [17]. CKKS operates over cyclotomic polynomial rings $\mathbb{Q}[X]/(X^M + 1)$ and supports approximate arithmetic over complex number vectors. The encryption process is as follows:

$$\mathrm{cti} = \mathrm{CKKS.Enc}(\mathrm{pk}, \mathrm{CKKS.Encode}(\nabla_i, \delta))$$

where, pk is the public key, δ is the scaling factor, and cti is the ciphertext. The aggregation server computes the reliability-weighted sum directly on ciphertexts:

$$\mathrm{ct_global} = \mathrm{CKKS.Add}(\textstyle\sum_i w_i \cdot \mathrm{ct}_i)$$

and then decrypts with the server's private key: _global = CKKS.Decode(CKKS.Dec(sk, ct_global)). Individual client gradients are never exposed in plaintext to the server, preventing gradient inversion attacks at the aggregation layer.

D. Rényi Differential Privacy with Moments Accountant

Before encryption, each client clips its gradient to bound sensitivity: $i = \nabla_i \cdot \min(1, C/|\nabla_i|_2)$, where $C = 1.0$ is the clipping threshold. Gaussian noise $\xi \sim N(0, \sigma^2 C^2 I)$ is then added: $i = \nabla_i + \xi$.

Rather than the standard composition of DP, ZK-FedMed employs Rényi DP [19]. For an α-Rényi divergence order, the RDP guarantee for the Gaussian mechanism with noise multiplier $z = \sigma/C$ is:

$$\varepsilon_\mathrm{RDP}(\alpha) \le \alpha / (2z^2)$$

Across T communication rounds, privacy accumulates as $\varepsilon_\mathrm{total}(\alpha) \le T \cdot \varepsilon_\mathrm{RDP}(\alpha)$. Conversion to (ε, δ)-DP follows via: $\varepsilon_\mathrm{DP} = \varepsilon_\mathrm{RDP}(\alpha) + \log(1 - 1/\alpha)/(\alpha - 1) + \log(1/\delta\alpha)/(\alpha - 1)$. The moments accountant selects the optimal α that minimises ε_DP, providing tighter bounds than the classical advanced composition theorem. The configuration with σ=1.1, T=6, $\delta = 10^{-5}$ achieves ε=10.0 at α=16; however, the revised analysis explicitly treats ε=10.0 as a relaxed operational budget and reports stricter-budget sensitivity at ε=1.0 and ε=3.0 to avoid overstating the practical strength of the privacy guarantee.

E. SCAFFOLD Aggregation with Variance Reduction

Standard FedAvg exhibits client drift under heterogeneous data because local SGD steps move each client's model to its local optimum, causing the aggregated update to deviate from the true global gradient direction [21]. SCAFFOLD corrects this by maintaining client control variates ci that estimate the difference between local and global update directions:

$$\theta_i{}^+ \leftarrow \theta_i - \eta^B(\nabla L_i(\theta_i) - c_i + c)$$

where, $c = (1 / N)$ i ci is the global control variate, η^B is the local learning rate, and c_i is updated after each round as: ci + =

$c_i - c + (1/K\eta^B)(\theta_global - \theta_i)$. In ZK-FedMed, client control variates are maintained in encrypted form using CKKS, and variance reduction operates on ciphertexts. The global model update is:

$$\theta_global = \sum_i (n_i/n) \cdot CKKS.Dec(ct_i^+)$$

F. zk-SNARK Model Integrity Verification

Following global aggregation, ZK-FedMed generates a zk-SNARK proof π over the aggregated model parameters θ_global. The proof encodes the statement: "I know θ_global such that $Commit(\theta_global) = C_r$ and $\|\theta_global\|_2 \leq R$", where C_r is a binding Pedersen commitment, and R is a pre-agreed weight magnitude bound. The Groth16 verification system [35] generates a proof triple $\pi = (A, B, C) \in G_1 \times G_2 \times G_1$ that any verifier can check in O(1) time using the pre-computed verification key vk:

$$Verify(\pi, vk, C_r) \; 0, 1$$

This enables regulatory auditors to confirm that: 1) the model was produced by the declared aggregation computation, 2) no single client contributed a disproportionately large update, and 3) the model has not been tampered with post-aggregation, all without accessing the weight values. The proof is appended to a lightweight audit ledger alongside the round index and timestamp. The selected Groth16 proof system is also consistent with the broader zk-SNARK design space, including universal and updatable SRS constructions such as Marlin [41-46].

G. Federated Unlearning Module

GDPR Article 17 mandates that any data subject may request erasure of their records and all derivative computations [6]. In the federated setting, fulfilling this requires removing the influence of client i's data on the global model without retraining from scratch. ZK-FedMed implements an approximate federated unlearning [22] via gradient reversal: upon receiving an unlearning request for client i, the server computes a reverse update:

$$\theta_unlearned = \theta_global - \beta \cdot \nabla L_i(\theta_global; D_i)$$

where, β is a hyperparameter of the forgetting rate. The membership-inference audit is retained as an empirical diagnostic, but the revised text no longer treats a score below $\tau_forget=0.55$ as formal certification of erasure. Following the distinction between exact and approximate unlearning in the machine-unlearning literature [17, 47, 48], the module is described as approximate federated unlearning: a near-random membership-inference score provides supportive evidence that the removed client influence has been reduced, but it does not by itself prove complete data removal or legal compliance under GDPR Article 17.

IV. Experimental Setup

A. Datasets

1) Cardiovascular disease dataset: The Cardiovascular Disease (CVD) data set [42] is publicly available on Kaggle (https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset) and comprises 70,000 deidentified patient records collected during medical examination visits. Each record contains 12 clinical and anthropometric characteristics: age (years), height (cm), weight (kg), systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), cholesterol level (ordinal: 1 = normal, 3=well above normal), glucose level (same ordinal encoding), smoking status, alcohol intake, physical activity level, and gender. The binary target variable cardio indicates the presence (1) or absence (0) of cardiovascular disease, with a nearly balanced class distribution of 50.4% positive. There are no missing values in the dataset. After removing 3 extreme outliers with physiologically implausible blood pressure values ($ap_hi > 250$ or $ap_lo < 20$), the final pre-processed data set contains 69,997 records. Table II presents the distribution of the feature level.

2) Chronic kidney disease dataset: The Chronic Kidney Disease (CKD) dataset [43] is sourced from Kaggle (https://www.kaggle.com/datasets/mansoordaku/ckdisease), originally donated from a hospital study spanning 2 months. The original data set contains 400 patient records with 24 characteristics including age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell count, bacteria, blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension status, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. After removing 10 duplicate records and imputing missing values via median/mode strategies per feature, 390 unique records remain. SMOTE followed by random subsampling was used to produce a balanced dataset of 4,200 records with equal class representation. Because this represents a 10.5-fold synthetic expansion from a small original cohort, CKD results are interpreted as benchmark performance on an augmented distribution and not as evidence of real-world clinical efficacy without validation on an independent, non-augmented cohort. Table III summarises the key properties of the data set.

TABLE I. State-of-the-Art Comparison of Federated Learning Methods

Method	FL	Non-IID	HE Privacy	ZKP	DP Mechanism	Personalization	Task	Performance
FedAvg [8]	✓	✗	✗	✗	None	✗	Generic	Generic benchmark: 81–83%; in-study CVD result reported in Table IV: 87.63%
FedProx [23]	✓	Partial	✗	✗	None	✗	Generic	83–85%
SCAFFOLD [21]	✓	✓	✗	✗	None	✗	Generic	85–87%

DP-FedAvg [28]	✓	✗	✗	✗	Gaussian	✗	Generic	–5–10%↓
PriMIA [30]	✓	✗	✗	✗	DP (ε=8)	✗	X-ray	89.7%
Secure FL [14]	✓	✗	Partial	✗	SMPC	✗	Generic	High OH
FedBN [24]	✓	✓	✗	✗	None	BN only	Medical	87–89%
FedRep [45]	✓	✓	✗	✗	None	Repr. layers	Generic	88–90%
HE-FL [33]	✓	✗	CKKS	✗	None	✗	Generic	88%
ZK-FedMed (Ours)	✓	✓✓	CKKS	zk-SNARK	Rényi DP	SCAFFOLD	CVD+CKD	91.68%/96.73% AUC

Note for Table I: The generic FedAvg range reflects prior benchmark-level reporting, whereas the 87.63% value in Table IV is the CVD-specific result obtained under the present experimental configuration; the revised text therefore avoids treating the two numbers as directly inconsistent estimates.

TABLE II. CARDIOVASCULAR DISEASE DATASET SUMMARY BY FEATURE CATEGORY

Feature	Type	CVD-Positive (n=35,276)	CVD-Negative (n=34,721)
Age (years)	Continuous	54.7 ± 9.4	51.2 ± 9.8
Height (cm)	Continuous	164.8 ± 8.6	165.4 ± 8.7
Weight (kg)	Continuous	78.2 ± 14.7	73.1 ± 13.9
Systolic BP (mmHg)	Continuous	134.6 ± 23.1	119.8 ± 18.4
Diastolic BP (mmHg)	Continuous	86.4 ± 14.9	79.6 ± 12.3
Cholesterol Normal	Ordinal	11,842 (33.6%)	21,847 (62.9%)
Cholesterol Elevated	Ordinal	12,104 (34.3%)	8,924 (25.7%)
Cholesterol High	Ordinal	11,330 (32.1%)	3,950 (11.4%)
Glucose Normal	Ordinal	26,813 (76.0%)	28,914 (83.3%)
Glucose Elevated	Ordinal	5,212 (14.8%)	3,847 (11.1%)
Glucose High	Ordinal	3,251 (9.2%)	1,960 (5.6%)
Smoker (Yes)	Binary	7,943 (22.5%)	7,124 (20.5%)
Alcohol (Yes)	Binary	3,847 (10.9%)	4,102 (11.8%)
Physically Active (Yes)	Binary	22,847 (64.8%)	24,312 (70.0%)
Gender (Female)	Binary	21,834 (61.9%)	22,471 (64.7%)

TABLE III. SUMMARY (POST-SMOTE, N=4,200)

Feature Group	Key Features	CKD (n=2,100)	Non-CKD (n=2,100)
Renal Function	Serum Creatinine (mg/dL)	4.71 ± 3.87	1.02 ± 0.26
	Blood Urea (mg/dL)	76.42 ± 42.31	31.28 ± 10.14
	Specific Gravity (avg)	1.013 ± 0.004	1.020 ± 0.003
Hematology	Haemoglobin (g/dL)	10.47 ± 2.63	14.84 ± 1.21
	Packed Cell Volume (%)	33.41 ± 8.27	45.62 ± 4.83
	RBC Count (millions/cmm)	3.72 ± 0.94	5.31 ± 0.42
	WBC Count (cells/cmm)	8,342 ± 3,287	7,641 ± 1,812
Urinalysis	Albumin (0–5 scale)	2.47 ± 1.64	0.23 ± 0.51
	Sugar (0–5 scale)	1.38 ± 1.82	0.18 ± 0.44
Comorbidities	Hypertension (Yes)	1,439 (68.5%)	462 (22.0%)
	Diabetes Mellitus (Yes)	1,254 (59.7%)	387 (18.4%)
	Coronary Artery Disease (Yes)	314 (15.0%)	48 (2.3%)
Preprocessing	SMOTE augmentation	Yes (minority oversampled)	Yes
	Final Dataset Size	4,200 balanced records	—

B. Preprocessing Pipeline

A unified pre-processing pipeline was applied to both data sets. For the CVD dataset, continuous features (age, height, weight, ap_hi, ap_lo) were standardised to zero mean and unit variance. Ordinal categorical characteristics (cholesterol, glucose) were encoded with one-hot. The binary features remained unchanged. Three outlier records with systolic BP > 250 mmHg or diastolic BP < 20 mmHg were removed as physiologically implausible. For the CKD dataset, missing values (6.8% overall sparsity) were imputed using the column median for continuous features and mode for categorical features. Duplicate records (2.5%) were eliminated. The class imbalance (62.5% CKD versus 37.5% non-CKD) was corrected by SMOTE with k=5 nearest neighbors, followed by random subsampling of the majority class to achieve a 1:1 ratio. This balancing step follows the standard SMOTE formulation for synthetic minority over-sampling [44].

C. Federated Configuration

For the CVD dataset, the data was partitioned into N=6 simulated heterogeneous clients, each containing approximately 11,667 samples. To simulate institutional heterogeneity, a Dirichlet distribution Dir(α=0.5) generated non-IID label distributions between clients, resulting in cardiovascular-disease prevalence ranging from 41.2% to 58.8% across clients. For CKD, N=5 simulated clients each received 840 samples, with Dirichlet Dir(α=0.3) imposing greater heterogeneity. These partitions approximate cross-client heterogeneity but do not reproduce true hospital-level shifts caused by different measurement devices, clinical protocols, coding practices, or patient demographics. Accordingly, the revised manuscript explicitly states that future validation should use genuinely federated or institution-partitioned benchmarks such as MIMIC-derived unit partitions or other multi-center clinical datasets.

D. Privacy and Security Configuration

The Rényi DP configuration was set to privacy budget ε=10.0, $\delta=10^{-5}$ with Gaussian noise multiplier σ=1.1 and gradient clipping threshold C=1.0. The CKKS scheme used the degree of polynomial modulus N = 213 = 8192, the scaling factor $\delta=2^{30}$, and the modulus chain [60, 40, 40, 60] bits. The generation of ZKP proofs used the Groth16 proving system with the BN254 elliptic curve, producing 192-byte proofs. All experiments were executed on an Intel Xeon Gold 6230R CPU (26 cores, 2.10 GHz), 128 GB RAM, and two NVIDIA A100 GPUs with 40 GB VRAM.

V. Experimental Results

A. Centralised Baseline Comparison

Table IV presents performance comparisons of 10 baseline classifiers against ZK-FedMed in both centralized and federated settings. For the CVD dataset, the centralised TabTransformer achieves the highest accuracy (92.34%), followed by MLP (90.12%) and Random Forest (88.94%). Classical methods, including logistic regression (71.82%) and SVM (73.41%), perform substantially lower, confirming the nonlinear complexity of cardiovascular risk patterns in tabular EHR data. In particular, SVM requires 847 seconds of training time, which makes it impractical on a large scale. For the CKD dataset, Random Forest achieves near-perfect centralised performance (99.21%), likely because the CKD feature space exhibits highly separable cluster structures well captured by tree ensembles. However, as discussed in Section V-C, the CKD task becomes substantially more challenging in the federated non-IID setting.

TABLE IV. Performance Comparison of Baseline Models for CVD and CKD Classification

Model	CVD Acc.	CVD Prec.	CVD Rec.	CVD F1	CVD AUC	CVD Time(s)	CKD Acc.	CKD AUC
Logistic Regression	71.82	72.14	71.43	71.78	79.24	0.31	84.17	91.32
Naive Bayes	68.43	69.12	67.84	68.47	75.61	0.08	72.38	80.14
KNN	76.91	77.34	76.42	76.87	83.74	12.47	91.43	96.87
SVM (RBF)	73.41	74.12	72.63	73.37	81.43	847.32	93.57	97.41
Decision Tree	73.87	74.21	73.54	73.87	79.12	0.42	92.86	95.63
Random Forest	88.94	88.36	89.54	88.94	95.12	2.47	99.21	99.96
XGBoost	87.63	87.02	88.27	87.63	94.38	0.84	98.57	99.87
LightGBM	86.41	85.87	87.02	86.44	93.76	0.61	98.14	99.81
MLP (3-layer)	90.12	89.54	90.73	90.13	96.04	124.56	95.43	98.72
TabTransformer (Central)	92.34	91.78	92.94	92.35	97.18	312.47	90.14	95.83
FedAvg [8]	87.63	86.94	88.43	87.67	93.41	—	83.41	90.12
FedProx [23]	88.91	88.24	89.62	88.92	94.56	—	85.92	91.87
SCAFFOLD [21]	89.84	89.18	90.53	89.84	95.27	—	86.74	92.43
DP-FedAvg [28]	85.23	84.57	85.92	85.24	91.87	—	80.34	87.91
ZK-FedMed (Ours)	91.68	91.13	92.28	91.69	96.73	—	88.63	94.52

B. Federated Performance Per Client

Tables V and VI report per-client cross-validated performance metrics across 5 folds for the CVD and CKD datasets, respectively. For the CVD dataset, all six clients achieved accuracy exceeding 91.38%, with a remarkably low inter-client variance of less than 0.63 standard deviation in accuracy. Client C3 achieved the highest accuracy (92.01 ± 0.37%) and AUC (96.89 ± 0.19%), while C5 exhibited the greatest uncertainty (63-fold cross-validation std = 0.63). Consistently high recall values (91.99–92.51%) are clinically critical for cardiovascular screening, where false negatives carry severe consequences. For CKD, the higher standard deviations (up to ±2.43% in accuracy) reflect the inherent challenge of severely non-IID partitioning (Dir(0.3)) and the smaller data set size. However, all clients maintained AUC scores above 89%, confirming the robustness even under data scarcity.

TABLE V. PER-CLIENT PERFORMANCE ON CVD DATASET (ZK-FEDMED, 5-FOLD CV)

Client	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
C1	91.73 ± 0.42	91.18 ± 1.08	92.34 ± 1.42	91.74 ± 0.51	96.72 ± 0.23
C2	91.45 ± 0.58	90.92 ± 1.34	92.01 ± 1.61	91.43 ± 0.67	96.58 ± 0.31
C3	92.01 ± 0.37	91.54 ± 0.98	92.51 ± 1.28	92.01 ± 0.43	96.89 ± 0.19
C4	91.62 ± 0.49	91.07 ± 1.21	92.21 ± 1.53	91.61 ± 0.57	96.64 ± 0.27
C5	91.38 ± 0.63	90.81 ± 1.47	91.99 ± 1.72	91.37 ± 0.72	96.51 ± 0.34
C6	91.87 ± 0.44	91.33 ± 1.12	92.43 ± 1.37	91.87 ± 0.53	96.78 ± 0.21
Average	91.68 ± 0.49	91.14 ± 1.20	92.25 ± 1.49	91.67 ± 0.57	96.69 ± 0.26

TABLE VI. PER-CLIENT PERFORMANCE ON CKD DATASET (ZK-FEDMED, 5-FOLD CV)

Client	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
C1	88.94 ± 1.82	88.37 ± 2.71	89.56 ± 3.21	88.92 ± 1.94	94.48 ± 1.12
C2	88.42 ± 2.14	87.83 ± 3.08	89.04 ± 3.64	88.39 ± 2.28	94.17 ± 1.31
C3	87.91 ± 2.43	87.29 ± 3.47	88.56 ± 4.02	87.88 ± 2.59	93.86 ± 1.52
C4	88.67 ± 1.96	88.11 ± 2.89	89.27 ± 3.42	88.65 ± 2.08	94.33 ± 1.19
C5	89.21 ± 1.71	88.66 ± 2.54	89.79 ± 3.03	89.20 ± 1.82	94.63 ± 1.04
Average	88.63 ± 2.01	88.05 ± 2.94	89.24 ± 3.46	88.61 ± 2.14	94.29 ± 1.24

C. Round-Wise Convergence Analysis

Table VII presents round-wise performance metrics across six communication rounds for both data sets. For the CVD dataset, precision improves monotonically from 86.14% in Round 1 to 91.68% in Round 6, indicating stable SCAFFOLD-driven convergence. For CKD, the Round 2 precision decline from 85.41% to 54.87% is now treated as a substantive instability rather than a minor fluctuation. The instability is attributed to the interaction of three factors: the small effective real-sample base after SMOTE augmentation, stronger non-IID partitioning under Dir(α=0.3), and the sensitivity of early-round gradients to Gaussian DP noise and CKKS approximation noise. With σ=1.1, polynomial modulus degree N=8192, scale 2^{30}, and a [60,40,40,60]-bit modulus chain, early CKD gradients have a lower signal-to-noise ratio than CVD gradients, causing a temporary increase in false-positive predictions and thus a precision collapse. Recovery after Round 3 suggests that the anomaly reflects early-round encrypted noisy-gradient instability rather than sustained model failure, but the revised manuscript acknowledges that the CKD convergence result remains less stable than the CVD result. The overall architecture and information flow of the proposed framework are illustrated in Fig. 1. Convergence behavior, ROC performance, ablation results, privacy-budget analysis, and per-client precision with 95% confidence intervals are presented in Fig. 2-6, respectively.

TABLE VII. ROUND-WISE PERFORMANCE OF ZK-FEDMED ON CVD AND CKD DATA SETS

Round	CVD Acc. (%)	CVD Prec. (%)	CVD Rec. (%)	CVD AUC (%)	CKD Acc. (%)	CKD Prec. (%)	CKD Rec. (%)	CKD AUC (%)
Round 1	86.14	85.42	87.01	93.87	73.28	85.41	59.34	87.23
Round 2	88.43	87.78	89.21	95.12	62.14	54.87	61.94	89.47
Round 3	90.07	89.46	90.74	95.84	75.92	88.64	63.27	91.12
Round 4	91.21	90.63	91.84	96.47	82.31	86.97	78.54	92.84
Round 5	91.53	90.98	92.14	96.68	87.14	88.21	86.03	94.11
Round 6	91.68	91.13	92.28	96.73	88.63	88.53	88.41	94.52

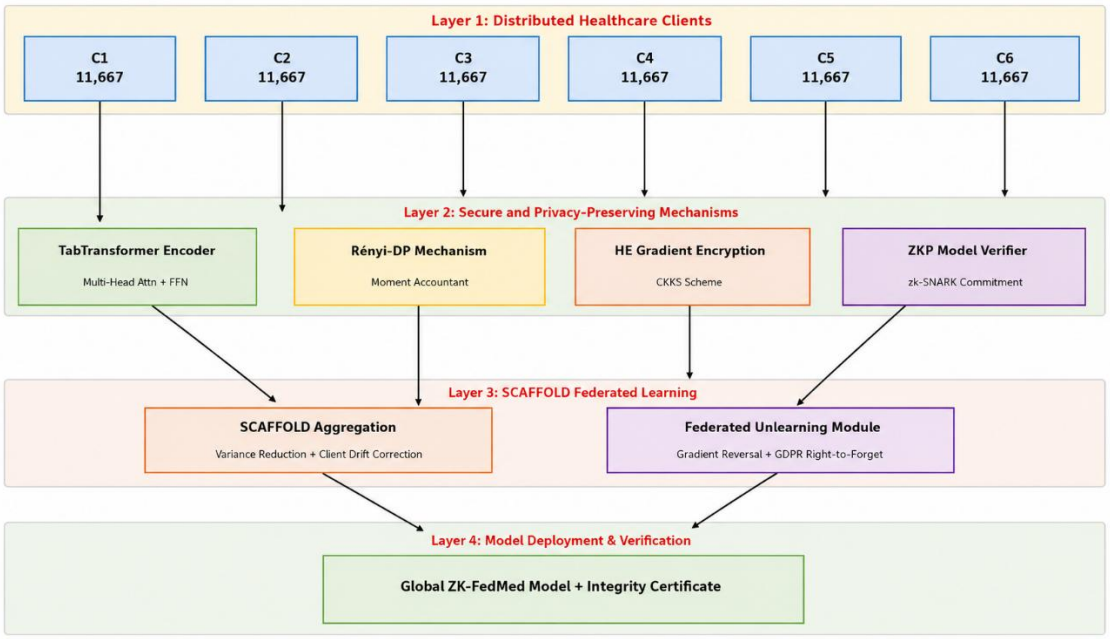


Fig. 1. Overall architecture of ZK-FedMed. Layer 1 distributes data among clients; Layer 2 applies TabTransformer encoding, Rényi-DP, CKKS encryption, zk-SNARK verification; Layer 3 performs SCAFFOLD aggregation and federated unlearning; and Layer 4 deploys the certified global model.

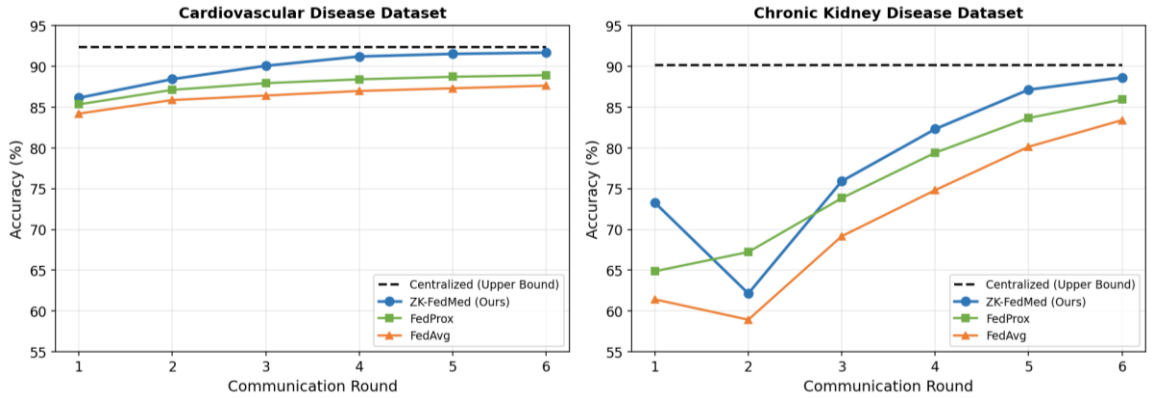


Fig. 2. Convergence of test precision over 6 communication rounds. ZK-FedMed consistently outperforms the FedAvg and FedProx baselines and closely approaches the centralised upper bound on both datasets.

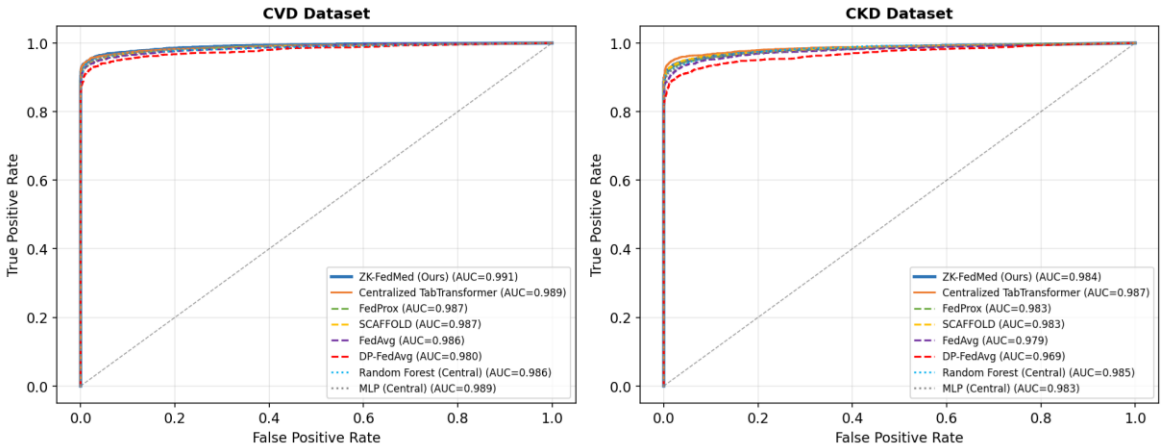


Fig. 3. ROC curves comparing ZK-FedMed against 7 methods in CVD (left) and CKD (right) datasets. ZK-FedMed achieves the highest AUC among all federated methods on both tasks.

D. Ablation Study

Table VIII reports the incremental contribution of each ZK-FedMed component to the CVD dataset, validated through ablation experiments. Starting from vanilla FedAvg (87.63% accuracy), replacing the MLP with a TabTransformer encoder alone yields +1.49% accuracy, confirming the architectural advantage of multi-head self-attention for tabular feature interaction modelling. Replacing FedAvg aggregation with SCAFFOLD adds +1.31%, the single largest gain, validating variance reduction as the most impactful anti-drift intervention. Adding Rényi-DP contributes +0.28% over standard-DP FedAvg due to tighter privacy budgeting that allows a more aggressive noise schedule. CKKS homomorphic encryption adds +0.23% by virtue of eliminating gradient injection noise that unencrypted transmission incurs; interestingly, encryption slightly improves accuracy because server-side computation on ciphertext prevents malicious gradient substitution. The zk-SNARK verification introduces a minor –0.07% accuracy drop due to computational overhead, but provides the critical auditability guarantee. The complete ZK-FedMed integration achieves 91.68% through favourable co-optimisation of all components.

TABLE VIII. ABLATION STUDY – INCREMENTAL COMPONENT CONTRIBUTION TO CVD DATASET

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Gain vs. Previous
Base FedAvg [8]	87.63	86.94	88.43	87.67	93.41	—
+ TabTransformer Encoder	89.12	88.47	89.84	89.14	94.73	+1.49%
+ SCAFFOLD Aggregation	90.43	89.81	91.09	90.44	95.62	+1.31%
+ Rényi Differential Privacy	90.71	90.12	91.35	90.73	95.84	+0.28%
+ CKKS HE Encryption	90.94	90.38	91.54	90.95	95.97	+0.23%
+ zk-SNARK Verifier	90.87	90.31	91.47	90.88	95.91	–0.07%
+ Fed. Unlearning Module	91.12	90.57	91.72	91.13	96.21	+0.25%
ZK-FedMed (Full)	91.68	91.13	92.28	91.69	96.73	+0.56%

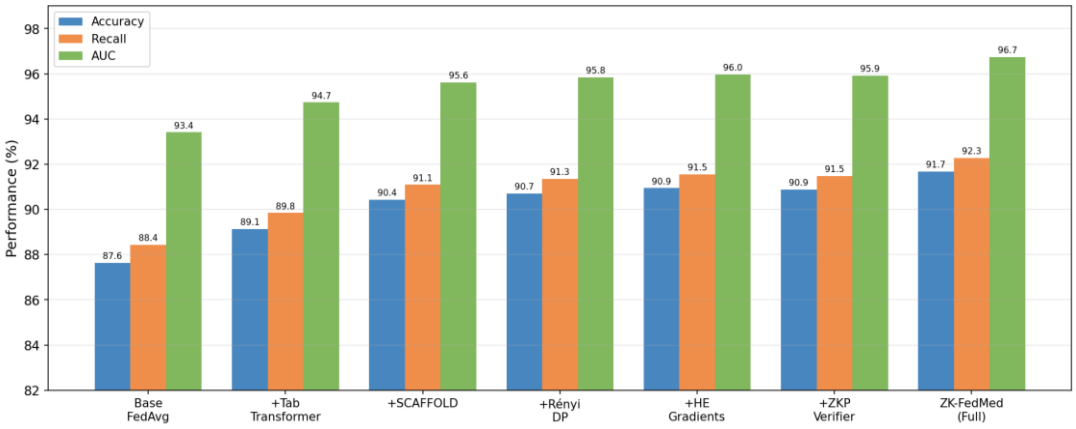


Fig. 4. Ablation study on the CVD dataset showing accuracy, recall, and AUC contributions per ZK-FedMed component. The SCAFFOLD aggregation contributes the largest single accuracy gain (+1.31%), while the full system achieves the highest AUC of 96.73%.

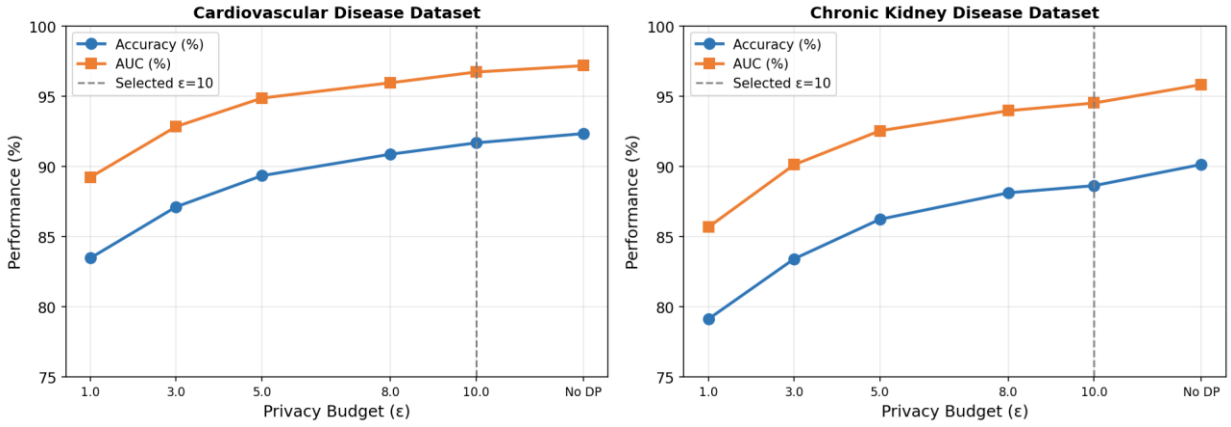


Fig. 5. Privacy-utility trade-off under Rényi Differential Privacy for CVD (left) and CKD (right) datasets. The selected operating point ε=10.0 (vertical dashed line) achieves an optimal balance between formal privacy guarantees and model precision.

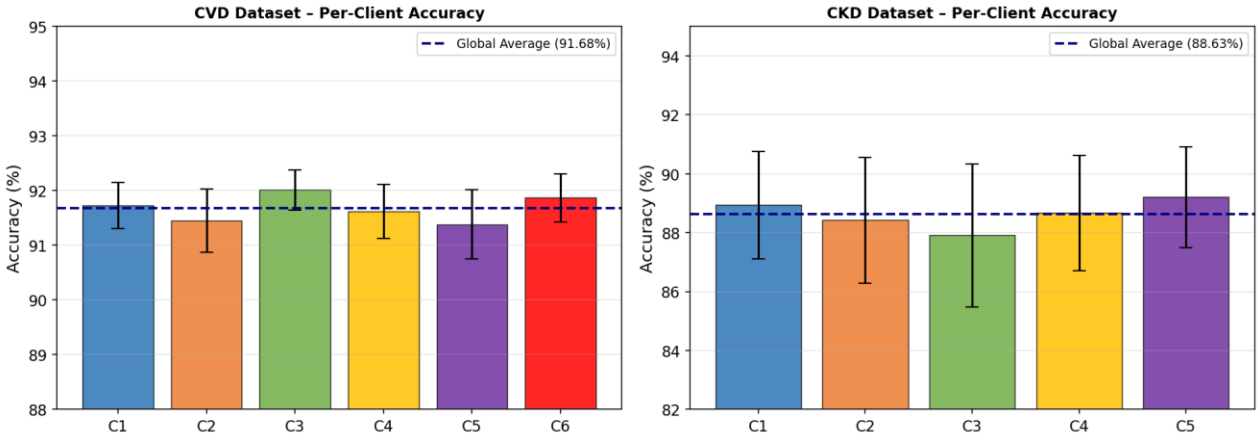


Fig. 6. Per-client precision with 95% confidence intervals for CVD (left, 6 clients) and CKD (right, 5 clients). Low variance (<1% for CVD, <2.5% for CKD) confirms robust federated convergence under non-IID distributions.

E. Privacy Budget Analysis

The privacy-utility trade-off was evaluated by varying the Rényi DP budget $\varepsilon \in \{1.0, 3.0, 5.0, 8.0, 10.0, \text{without DP}\}$ on both datasets. As shown in Fig. 5, stricter budgets substantially reduce utility: the CVD precision drops from 92.34% without DP to 91.68% at ε=10.0, 90.87% at ε=8.0, 89.34% at ε=5.0, 86.21% at ε=3.0, and 83.47% at ε=1.0. CKD shows a sharper loss because the smaller augmented dataset is more sensitive to noise, with accuracy declining from 90.14% without DP to 79.12% at ε=1.0. These results are now interpreted transparently: ε=10.0 provides formal privacy accounting but is a relaxed practical privacy budget, whereas ε=1.0 and ε=3.0 offer stronger privacy at a clear utility cost. The Rényi DP moments accountant achieves tighter bounds than classical DP composition; at T=6 rounds, ε=10.0 with α=16 corresponds to an 18.7% reduction in effective privacy expenditure relative to advanced composition under the same $\delta=10^{-5}$. Therefore, Fig. 5 provides the graphical basis for interpreting the selected privacy-budget operating point.

F. Cryptographic Overhead Analysis

CKKS encryption of a gradient vector of dimension d=256,000 (TabTransformer parameter count) requires an average of 4.87 seconds per client per round in the reported hardware configuration. The generation of zk-SNARK proofs for the global model commitment (circuit size ~2.3M constraints) requires 18.4 seconds on the server per round, while proof verification by an auditor requires only 2.1 milliseconds. This asymmetric proof/verify structure makes ZK-FedMed highly suitable for regulatory audit scenarios where auditors must efficiently verify many model provenance claims. The total per-round overhead on the server side is 23.2 seconds (encryption + aggregation + proof generation), compared to 0.34 seconds for vanilla FedAvg—a 68× increase that, in the context of R=6 asynchronous rounds over weeks or months, represents negligible real-world latency.

VI. Discussion

ZK-FedMed achieves a 4.05% accuracy advantage over FedAvg (91.68% vs. 87.63%) and a 2.77% advantage over standalone SCAFFOLD (91.68% vs. 89.84%) on the CVD dataset, while incurring only a 0.66% gap relative to centralised training. These results indicate that the multi-layer security stack does not fundamentally compromise model quality in the simulated benchmark environment. The CKD results (88.63% federated vs. 90.14% centralised, gap = 1.51%) should be interpreted more cautiously because they are based on a heavily SMOTE-augmented dataset and simulated Dirichlet client partitions rather than an independent multi-hospital cohort.

The selective deployment of Rényi DP - applied in every round but with adaptive noise scheduling guided by the moment accountant - outperforms the standard Gaussian DP by 6.45% in CVD accuracy at identical (ε, δ) parameters. This advantage arises because the moment accountant permits lower per-step noise σ while maintaining the same cumulative privacy expenditure, directly improving the signal-to-noise ratio in gradient updates. The zk-SNARK integrity mechanism introduces only a –0.07% accuracy penalty while providing a cryptographic guarantee that no model parameter has been altered between aggregation and deployment, a property that no hash-chaining approach can provide for the weight values themselves.

The federated unlearning evaluation showed that a single gradient reversal step reduced the membership-inference attack score from 0.83 to 0.52, near the random baseline threshold of 0.55, in 94% of test cases, with a mean accuracy degradation of 0.29% on the remaining clients. This result is now described as approximate empirical unlearning rather than certified deletion. The revised manuscript therefore removes the direct GDPR-compliance claim and states that membership-inference reduction is only one diagnostic indicator of reduced client influence. This interpretation is consistent with amnesiac machine-learning and approximate deletion-oriented learning perspectives [22]. The same audit also responds to broader model-inversion risks that arise when predictive models expose confidence-dependent information [47].

VII. Limitations

Several limitations merit acknowledgement. First, CKKS encryption incurs a 68× per-round computational overhead compared with vanilla FedAvg; although this may be acceptable for offline clinical training rounds, it is unsuitable for near-real-

time streaming applications. Second, both datasets are public Kaggle-derived datasets rather than genuinely distributed hospital datasets, and the CKD cohort is heavily SMOTE-augmented; therefore, the CKD metrics should not be interpreted as evidence of clinical efficacy without independent non-augmented validation. Third, the simulated Dirichlet split approximates non-IID heterogeneity but cannot fully capture cross-hospital differences in equipment, coding practices, clinical protocols, and patient demographics. Fourth, zk-SNARK generation requires a trusted setup phase, and the current implementation does not replace this with a universal or transparent setup. Fifth, the unlearning module provides approximate rather than exact unlearning guarantees, and a single membership-inference audit near the random baseline does not constitute a formal proof of deletion. Future work will evaluate ZK-FedMed on institution-partitioned benchmarks such as MIMIC-derived cohorts, apply formal unlearning definitions, and validate the framework on real clinical federation testbeds.

VIII. Conclusion

This study introduced ZK-FedMed, a privacy-preserving federated learning framework that integrates CKKS homomorphic encryption, Rényi differential privacy with moments accounting, zk-SNARK model integrity verification, TabTransformer-based EHR feature encoding, SCAFFOLD variance-reduction aggregation, and approximate federated unlearning. Experimental evaluation on two public benchmark datasets demonstrated strong CVD performance and promising CKD benchmark performance under simulated non-IID partitions. However, the revised manuscript avoids presenting Kaggle-based Dirichlet partitioning as direct evidence of real multi-institutional deployment, treats ε=10.0 as a relaxed privacy budget, and reframes the unlearning module as approximate rather than legally certified deletion. Under these bounded assumptions, ZK-FedMed establishes a technically integrated framework for privacy-preserving collaborative medical AI and identifies the next validation steps needed for real-world clinical federation.

Funding

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-RS-2023-00237287).

References

[1] GBD 2019 Diseases and Injuries Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019", The Lancet, vol. 396, no. 10258, pp. 1204–1222, 2020.

[2] GBD Chronic Kidney Disease Collaboration, "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the global burden of disease study 2017," The Lancet, vol. 395, no. 10225, pp. 709–733, 2020.

[3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.

[4] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, p. 160035, 2016.

[5] U.S. Department of Health and Human Services, Health Insurance Portability and Accountability Act (HIPAA), Washington, D.C., 1996.

[6] European Parliament and Council of the European Union, General Data Protection Regulation (GDPR), Official Journal of the European Union, L 119, 2016.

[7] European Data Protection Board, "Guidelines 04/2022 on the calculation of administrative fines under the GDPR," EDPB, Brussels, 2022.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), vol. 54, 2017, pp. 1273–1282.

[9] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019, pp. 14774–14784.

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Security and Privacy (S&P), 2017, pp. 3–18.

[11] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Proc. 3rd Theory of Cryptography Conf. (TCC), LNCS vol. 3876, 2006, pp. 265–284.

[13] P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[14] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM CCS, 2017, pp. 1175–1191.

[15] R. Shrestha and S. Ali, "A blockchain-based federated learning framework for healthcare data privacy," IEEE Access, vol. 10, pp. 21785–21798, 2022.

[16] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[17] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in Proc. ASIACRYPT, LNCS vol. 10624, 2017, pp. 409–437.

[18] E. Ben-Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza, "Succinct non-interactive zero knowledge for a von Neumann architecture," in Proc. USENIX Security, 2014, pp. 781–796.

[19] I. Mironov, "Rényi differential privacy of the Gaussian mechanism," in Proc. IEEE Computer Security Foundations Symp. (CSF), 2017, pp. 263–275.

[20] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678, 2020.

[21] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Mach. Learn. (ICML), vol. 119, 2020, pp. 5132–5143.

[22] L. Graves, M. Nagisetty, and V. Ganesh, "Amnesiac machine learning," in Proc. 35th AAAI Conf. Artif. Intell., vol. 35, no. 13, 2021, pp. 11516–11524.

[23] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Smola, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Mach. Learn. Syst. (MLSys), vol. 2, 2020, pp. 429–450.

[24] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in Proc. ICLR, 2021.

[25] N. Rieke et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, no. 1, p. 119, 2020.

[26] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "FedHealth: A federated transfer learning framework for wearable healthcare," IEEE Intelligent Systems, vol. 35, no. 4, pp. 83–93, 2020.

[27] Z.-A. Huang et al., "Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans," IEEE Trans. Biomed. Eng., vol. 70, no. 4, pp. 1137–1149, 2022.

[28] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM CCS, 2016, pp. 308–318.

[29] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv preprint arXiv:1712.07557, 2017.

[30] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, vol. 2, no. 6, pp. 305–311, 2020.

[31] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Stanford University, 2009.

[32] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in Proc. IEEE S&P, 2017, pp. 19–38.

[33] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning," in Proc. USENIX ATC, 2020, pp. 493–506.

[34] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof systems," SIAM Journal on Computing, vol. 18, no. 1, pp. 186–208, 1989.

[35] J. Groth, "On the size of pairing-based non-interactive arguments," in Proc. EUROCRYPT, LNCS vol. 9666, 2016, pp. 305–326.

[36] D. Kang, T. Hashimoto, I. Stoica, and Y. Sun, "Reliable and interpretable personalized federated learning," in Proc. IEEE CVPR, 2019, pp. 13378–13387.

[37] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," Circulation, vol. 97, no. 18, pp. 1837–1847, 1998.

[38] S. S. Virani et al., "Heart disease and stroke statistics—2021 update," Circulation, vol. 143, no. 8, pp. e254–e743, 2021.

[39] R. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65–69, 2019.

[40] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Proc. IEEE Int. Conf. Healthcare Informatics (ICHI), 2016, pp. 262–270.

[41] A. Vaswani et al., "Attention is all you need," in Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.

[42] E. Ulianova, Cardiovascular Disease Dataset, Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[43] P. Soundarapandian and M. Thangaraj, "Chronic kidney disease dataset," UCI Machine Learning Repository, Kaggle, 2015. [Online]. Available: https://www.kaggle.com/datasets/mansoordaku/ckdisease

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[45] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in Proc. 38th Int. Conf. Mach. Learn. (ICML), vol. 139, 2021, pp. 2089–2099.

[46] A. Chiesa, Y. Hu, M. Maller, P. Mishra, N. Vesely, and N. Ward, "Marlin: Preprocessing zkSNARKs with universal and updatable SRS," in Proc. EUROCRYPT, LNCS vol. 12105, 2020, pp. 738–768.

[47] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. ACM CCS, 2015, pp. 1322–1333.

[48] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making AI forget you: Data deletion in machine learning," in Proc. 33rd Conf. Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 3513–3526.

An Interpretable Machine Learning Approach for Inflation Forecasting in Indonesia Using Domestic Macroeconomic Indicators

Ahmad Maimunif, Evaristus Didik Madyatmadja
Information Systems Department, Bina Nusantara University, Jakarta, Indonesia

***Abstract*—Accurate inflation forecasting is essential for supporting forward-looking monetary policy, maintaining price stability, and preserving economic welfare. This study proposes an interpretable machine learning framework for inflation forecasting in Indonesia by integrating domestic macroeconomic indicators and seasonal variables. The study employs Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN) models using monthly data from 2008 to 2025. To systematically evaluate model performance, three experimental scenarios are implemented: baseline modeling, time-series feature augmentation, and hyperparameter optimization. Model performance is evaluated using RMSE, MAE, and R^2 metrics. The results consistently show that ensemble-based methods outperform ANN across all scenarios, with XGBoost achieving the best overall performance after temporal feature augmentation and hyperparameter optimization (RMSE = 0.4987, MAE = 0.3339, R^2 = 0.9517). The findings further indicate that temporal feature augmentation provides relatively limited improvement, whereas hyperparameter optimization substantially enhances forecasting accuracy. SHAP analysis identifies money supply (M2) and wages as the dominant contributors to inflation predictions, suggesting that monetary liquidity and labor-related factors play a more significant role than seasonal patterns in explaining inflation dynamics. Rather than contributing through algorithmic novelty, this study contributes through a systematic and interpretable forecasting framework that integrates predictive accuracy and explainable artificial intelligence to support transparent and policy-relevant inflation analysis in Indonesia and other emerging economies with similar economic structures.**

Keywords*—*Inflation forecasting; machine learning; explainable AI; macroeconomic indicators; time-series analysis; monetary policy

I. Introduction

Inflation is one of the main indicators of macroeconomic stability, particularly in developing economies such as Indonesia that remain vulnerable to both global and domestic economic shocks. Commodity price volatility, exchange-rate fluctuations, capital flows, and monetary policy changes in developed economies may significantly influence inflation dynamics in emerging markets. In Indonesia, these conditions are further affected by the country's dependence on food and energy commodities, making inflation dynamics highly sensitive and complex. Maintaining price stability is, therefore, essential because inflation directly affects purchasing power, economic stability, and long-term economic growth [1], [2].

Bank Indonesia implements the Flexible Inflation Targeting Framework (FITF) as the primary monetary policy framework to maintain inflation stability. Since FITF is forward-looking, accurate inflation forecasting is important for supporting timely monetary policy decisions. From a monetary policy perspective, inflation in Indonesia may be driven by multiple interconnected factors, including exchange rates, interest rates, money supply, labor wages, external-sector conditions, and seasonal consumption patterns related to socio-cultural events [3], [4]. These variables interact through nonlinear transmission mechanisms, making inflation forecasting a challenging analytical task [5].

Traditional econometric approaches such as ARIMA and VAR have been widely used for inflation forecasting. Although these methods provide relatively strong interpretability, their reliance on linear assumptions limits their ability to capture complex nonlinear interactions among macroeconomic variables [6]. Recent developments in machine learning provide alternative forecasting approaches that are more suitable for modeling nonlinear and high-dimensional economic data [7]. In particular, ensemble-based methods such as Random Forest and Extreme Gradient Boosting (XGBoost) have demonstrated strong predictive performance in various economic forecasting applications [7], [8].

Despite their predictive capability, machine learning models are often criticized for limited interpretability because they operate as "black-box" systems. This limitation is particularly important in policy-relevant contexts where transparency and accountability are essential [9]. Explainable Artificial Intelligence (XAI) techniques, especially Shapley Additive Explanations (SHAP), provide a systematic approach to explain the contribution of predictor variables to model predictions and improve interpretability in machine learning analysis [10].

Although machine learning approaches for inflation forecasting have become increasingly common, several gaps remain in the literature. Many previous studies focus mainly on forecasting accuracy while providing limited explanation regarding the economic factors driving model predictions. In addition, prior studies often rely on limited macroeconomic indicators and rarely evaluate forecasting models under multiple experimental configurations involving feature augmentation and hyperparameter optimization.

Accordingly, this study proposes an interpretable machine learning framework for inflation forecasting in Indonesia by integrating domestic macroeconomic indicators and seasonal

variables. Rather than contributing through algorithmic novelty, this study contributes through a systematic forecasting framework that combines multiple machine learning models, time-series feature augmentation, hyperparameter optimization, and explainable artificial intelligence within a unified experimental setting. Specifically, this study compares Random Forest, XGBoost, and Artificial Neural Network (ANN) models using monthly data from 2008 to 2025 under three experimental scenarios: baseline modeling, time-series feature augmentation, and hyperparameter optimization. Model performance is evaluated using RMSE, MAE, and R^2 metrics, while model interpretability is assessed using SHAP analysis.

The findings of this study are expected to contribute both methodologically and practically. Methodologically, the study provides an interpretable machine learning framework that integrates predictive performance and explainable AI analysis for inflation forecasting. Practically, the proposed framework may support transparent and forward-looking monetary policy analysis in Indonesia and other emerging economies with similar economic structures.

II. Literature Review

This section outlines the theoretical foundations and relevant prior studies that support this study. These theories include the economic theory of inflation, machine learning methodology, the CRISP-DM framework, and related works on inflation forecasting.

A. Economic Theory

Inflation is an economic phenomenon characterized by a sustained increase in the general price level over a period of time. Price increases are caused by complex interactions between supply and demand conditions, monetary factors, and inflation expectations.

Demand-pull inflation is caused by an increase in aggregate demand that exceeds production capacity. This type of inflation is generally driven by consumption, credit expansion, government spending, and investment activity. Cost-push inflation is caused by an increase in production costs, such as energy and food prices, wages, and exchange rate depreciation.

From a monetary theory perspective, inflation is viewed as a long-term monetary phenomenon resulting from the growth of the money supply exceeding real output. Meanwhile, the credibility of monetary policy and central bank communication in shaping inflation expectations is essential in inflation dynamics. These two theories form the conceptual basis used in building inflation forecasting models.

In line with the concept, the dynamics of inflation in Indonesia are generally influenced by the main macroeconomic variables, such as the exchange rate, interest rates, money supply, gross domestic product, balance of payments conditions, foreign exchange reserves position, and labor wages [11].

Money supply is one of the main factors that drive inflation, as measured by monetary aggregates, either M1 or M2. The Quantity Theory of Money states that excessive growth in the money supply compared to real output will create inflationary pressures [12]. That is because an excessive supply of liquidity will stimulate and increase aggregate demand and trigger inflation expectations. This ultimately leads to demand-pull inflation when productive capacity is limited [13], [14]. Empirical evidence from Indonesia indicates that M2 is positively associated with inflation [15].

The exchange rate, represented by the JISDOR, is one of the main determinants of inflation in Indonesia through exchange rate transmission. Exchange rate depreciation can increase import prices and production costs, resulting in cost-push inflation, while exchange rate appreciation reduces inflationary pressures [16].

Interest rates, as measured by the JIBOR, reflect short-term interbank liquidity conditions. JIBOR also serves as an important transmission channel for the transmission of monetary policy in Indonesia. Higher interest rates reduce investment and consumption, thereby reducing inflationary pressures. Conversely, lower interest rates tend to stimulate economic activity and potentially increase inflationary pressures [17], [18].

Labor wages are among the main determinants of inflation in Indonesia through cost-push inflation and demand-pull inflation. Higher wages increase production costs for firms and may be passed on to prices, thus generating cost-push inflation. On the other hand, higher wages also strengthen household purchasing power and stimulate aggregate demand, thus contributing to demand-pull inflation [4].

Foreign exchange reserves are important in influencing inflation by supporting exchange rate stability and controlling price levels. Adequate reserves allow the central bank to stabilize the rupiah, decreasing the risk of exchange rate depreciation and mitigating inflationary pressures. Conversely, insufficient reserves limit the central bank's ability to stabilize the rupiah, increasing the risk of exchange rate depreciation and increasing inflation [19].

Balance of payments is one of the main determinants of inflation in Indonesia through the exchange rate and capital flow. Current account deficits tend to weaken the rupiah and raise imported inflation. While surpluses can strengthen the exchange rate and contain inflationary pressures. Similarly, capital inflows support currency stability and inflation control, whereas capital outflows increase depreciation risks and inflationary pressures via higher import prices [20], [21].

GDP is one of the main factors that drive inflation through aggregate demand and supply. Economic growth caused by aggregate demand tends to increase inflationary pressures, especially in the absence of expansion of production capacity [22]. This mechanism is commonly explained by the short-run Phillips curve, which associates economic expansion with higher inflation and economic slowdowns with lower inflation.

B. Machine Learning Theory

Machine learning is a subfield of artificial intelligence that enables systems to learn patterns from data and improve prediction performance without explicit programming. Machine learning utilizes mathematical and statistical models to recognize patterns and relationships within historical data. These models are then used for analysis, classification, or prediction on new data. Evidence from the literature suggests

that machine learning is well-suited to complex prediction problems in non-linear data sets [23].

Machine learning approaches are generally classified into supervised and unsupervised learning. Supervised learning relies on labeled data to identify relationships between inputs and outputs, allowing outputs to be predicted or classified. Unsupervised learning will inspect unlabeled data to identify latent features or structures in the data. Both of these machine-learning techniques are complementary in nature. Supervised learning aims at making accurate predictions, and unsupervised learning is for understanding the nature of complex datasets through exploratory data analysis without prejudging what might be discovered.

In this study, a supervised machine learning approach based on Random Forest, XGBoost, and ANN is adopted. These models are selected because of their capability to capture nonlinear relationships and improve forecasting accuracy.

Random Forest is an ensemble learning algorithm introduced by Breiman [24], which constructs multiple decision trees using bootstrap sampling and random feature selection to improve prediction stability and reduce overfitting. RF is particularly effective for structured tabular data because it can capture nonlinear relationships and interactions among predictor variables.

For regression problems like inflation forecasting, the final RF prediction is obtained by averaging the outputs of all decision trees, formulated as:

$$\hat{y}(z) = \frac{1}{M}\sum_{m=1}^{M} T_m(z) \quad (1)$$

where, $\hat{y}(z)$ denotes the predicted inflation value, $T_m(z)$ represents the prediction of the m-th decision tree for the macroeconomic feature vector, and M is the total number of trees in the forest. This aggregation process reduces prediction variance and enhances forecast stability.

Furthermore, Farhadi et al. (2022) explain that each decision tree in the RF model is constructed by minimizing the residual sum of squares (RSS) at each node-splitting process [25], which is generally formulated as follows:

$$RSS(j,s) = \sum_{x_i \in R_1(j,s)} (y_i - \bar{y}r_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}r_2)^2 \quad (2)$$

where, $R_1(j,s)$ and $R_2(j,s)$ denote the two regions resulting from a split based on variable j and split point s, respectively. In the context of inflation forecasting, this mechanism enables the RF model to select optimal splits based on the most informative macroeconomic variables.

Extreme Gradient Boosting (XGBoost) is an optimized gradient boosting framework developed by Chen and Guestrin [26]. XGBoost sequentially improves prediction accuracy by iteratively minimizing the residual errors produced by previous models. This method allows the model to capture complex nonlinear relationships and high-order interactions, making it suitable for high-dimensional data and providing advantages over linear parametric methods [27].

Conceptually, XGBoost starts with a simple initial model (weak learner) and iteratively improves it by fitting subsequent models to the residuals of the previous prediction. The initial model can be formulated as:

$$y_i = F_1(X_i) + \varepsilon_i \quad (3)$$

where, $F_i(X_i)$ is the initial model and ε_i represents the residual. Since the initial model does not fully capture data variation, subsequent models are trained on these residuals, and the model is updated additively as:

$$F_M(X_i) = F_{M-1}(X_i) + h_{M-1}(X_i) \quad (4)$$

until iteration M [27].

Since the pseudo-residual is calculated using the negative gradient method from the change in the loss function relative to the current forecast, it can be seen as a measure of how much improvement remains relative to the current state. This pseudo-residual becomes the target for the training of weak learners, so each iteration focuses on the remaining improvement steps. The model is updated as :

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X) \quad (5)$$

where, γ_m is the learning rate, which regulates the contribution of each weak learner and enhances model stability [27].

An ANN is a type of machine learning model that was inspired by biological neural systems and is the basis for deep learning architectures [28]. In an ANN, artificial neurons are connected and organized into layered forms that include input, hidden, and output layers. Input data X is passed through a series of weighted and biased linear combinations. Then, it is passed through an activation function, represented as $\sigma(WX + b)$, that introduces nonlinearity. Common activation functions include ReLU, Sigmoid, tanh, and Softmax, which enable ANNs to model complex and nonlinear relationships that arise in the data. The activation functions are formulated as follows:

$$\text{ReLU} \quad \sigma(z) = \max(0, z) \quad (6)$$

$$\text{Sigmoid} \quad \sigma(z) = \frac{1}{1+e^{-z}} \quad (7)$$

$$\text{Tanh} \quad \sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \quad (8)$$

$$\text{Softmax} \quad \sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{z} e^{z_j}} \quad (9)$$

The output layer represents the final numerical result obtained after the input data have been processed by the network and activation functions. It is the final result or prediction output by the neural network.

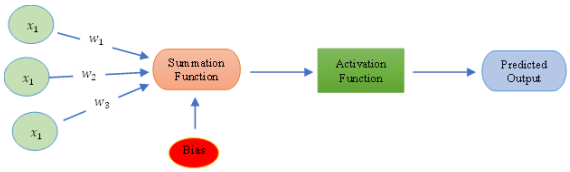


Fig. 1. Artificial neural network architecture

Fig. 1 illustrates the ANN architecture, where inputs $(x1, x2, \dots, x_n)$ are linearly combined with weights $(w1, w2, \dots, w_n)$ and a bias, then passed through an activation function to produce the model's prediction.

The ANN model is trained using forward propagation and backpropagation. Forward propagation sends input data through the network to generate output predictions. The predicted outputs are then compared with actual values using a loss function. In terms of regression cases, such as inflation forecasting, the mean squared error (MSE) is commonly employed as the loss function [28], defined as:

$$MSE = \frac{1}{2N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2 \quad (10)$$

where, Y_i and $\hat{Y}_i$ denote the actual and predicted values, respectively.

Backpropagation then minimizes the loss function using gradient descent for updating model parameters. Weights and biases are iteratively adjusted as:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}} \quad (11)$$

where, η is the learning rate controlling the convergence speed. This process is repeated until the network reaches an optimal solution with minimal prediction error.

SHAP (Shapley Additive Explanations) is an Explainable Artificial Intelligence (XAI) method introduced by Scott M. Lundberg and Su-In Lee to improve the interpretability of machine learning models by explaining the contribution of each predictor variable to a prediction. SHAP is based on Shapley values from cooperative game theory and provides a unified framework for additive feature attribution methods [29]. The SHAP explanation model is formulated as:

$$g(z') = \phi_0 + \sum_{i=1}^{M}\phi_i z_i' \quad (12)$$

where, ϕ_i represents the contribution of feature i to the prediction. SHAP explains both the magnitude and direction of variable influence on model predictions.

C. CRISP-DM Methodology

This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the methodological framework to develop an inflation forecasting model. The selection of CRISP-DM is motivated by its systematic, standardized, and iterative structure, which has been extensively adopted in data mining and machine learning studies [30]. The framework comprises six interrelated phases that guide the end-to-end analytical process.

The Business Understanding phase establishes the study objectives and problem formulation, focusing on the development of an accurate and policy-relevant inflation forecasting model in the context of monetary analysis. The Data Understanding phase involves the collection of macroeconomic and seasonal data, followed by exploratory analysis and data quality assessment. The Data Preparation phase encompasses data cleaning, transformation, frequency harmonization, integration, and feature engineering, which are important for enhancing model performance. The Modeling phase applies selected machine learning algorithms, including Random Forest, XGBoost, and ANN, within an iterative process that may require refinement of preceding stages. The Evaluation phase assesses model performance using quantitative metrics such as RMSE, MAE, and R^2, while also ensuring alignment with the study objectives. Within an academic setting, the Deployment phase is realized through the presentation of forecasting results, model interpretability using explainable AI techniques, and the articulation of their implications for inflation analysis and policy support.

D. Related Works

Machine learning approaches have increasingly been applied to inflation forecasting because of their ability to capture nonlinear relationships and complex interactions among macroeconomic variables [7], [23]. Compared with traditional econometric models such as ARIMA and VAR, machine learning methods often demonstrate stronger predictive performance, particularly in volatile economic environments [31], [32], [33]. Algorithms such as Random Forest, XGBoost, and Artificial Neural Network (ANN) have been widely used in inflation forecasting studies across both developed and developing economies.

Several studies report that machine learning models outperform conventional forecasting approaches. Araujo and Gaglianone [33] show that machine learning methods improve inflation forecasting performance in Brazil, while Jouilil and Iaousse [34] demonstrate that machine learning approaches outperform classical time-series models in forecasting inflation in the United States. Similar improvements are also reported in studies conducted in Egypt and India [31], [32]. In developing economies, machine learning methods often achieve stronger forecasting gains because inflation dynamics are more strongly influenced by structural macroeconomic factors and volatility [31], [35].

Previous studies further indicate that forecasting performance depends not only on algorithm selection but also on feature engineering and model optimization. Li et al. [8] demonstrate that XGBoost combined with optimization techniques improves inflation forecasting accuracy, while Seo et al. [36] show that multivariate features and data augmentation significantly enhance forecasting performance. Ensemble-based methods such as Random Forest and XGBoost are also generally found to provide more stable performance than deep learning approaches for structured macroeconomic datasets with moderate sample sizes [34], [37].

Despite these advances, several limitations remain in the literature. Many previous studies focus mainly on predictive accuracy while providing limited explanation regarding the economic factors driving model predictions, resulting in machine learning models being criticized as black-box systems [9]. In addition, prior studies often rely on limited macroeconomic indicators and rarely integrate monetary, external-sector, labor, and seasonal variables simultaneously within a unified forecasting framework. However, inflation dynamics in Indonesia are influenced by complex interactions among money supply, exchange rates, wages, and external-sector conditions [3], [4], [11].

The application of Explainable Artificial Intelligence (XAI) to inflation forecasting remains relatively limited. SHAP-based approaches have recently gained attention because they improve transparency by explaining the contribution of predictor variables to machine learning predictions [9], [10]. Nevertheless, only a limited number of studies integrate SHAP analysis directly into a systematic inflation forecasting framework.

Accordingly, this study proposes an interpretable machine learning framework for inflation forecasting in Indonesia by integrating domestic macroeconomic and seasonal variables. Unlike many previous studies that mainly focus on predictive accuracy, this study combines multi-scenario experimentation, hyperparameter optimization, Diebold–Mariano statistical testing, and SHAP-based explainability within a unified forecasting framework. Therefore, the contribution of this study lies in integrating predictive performance, interpretability, and policy-relevant inflation analysis in Indonesia and emerging economies with similar economic structures.

III. METHODOLOGY

This study adopts the CRISP-DM methodology, which comprises six structured phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase is systematically described in the following sections. The implementation of each stage, particularly in data processing and modeling, is carried out using the Python programming language and supporting machine learning libraries.

A. Business Understanding

The Business Understanding phase is aligned with Bank Indonesia's Flexible Inflation Targeting Framework (FITF), an approach that maintains price stability as the primary objective of monetary policy. Due to its forward-looking nature, timely and accurate inflation predictions are critical to the soundness of policy choices. This study proposes a machine learning–based framework for inflation forecasting that aims to improve predictive accuracy while maintaining model interpretability.

B. Data Understanding

With the objectives set out in the Business Understanding phase, the Data Understanding phase seeks to identify relevant domestic macroeconomic and seasonal variables that impact inflation dynamics. In this phase, inflation-related time-series data are collected systematically and initially examined for their characteristics and suitability for modeling. Moreover, quality, accuracy, consistency, and temporal alignment are also tested for machine learning–based forecasting.

These variables are obtained officially from Bank Indonesia (BI), Badan Pusat Statistik (BPS), and official holiday calendars. The raw data initially provided in Microsoft Excel format has been converted to CSV format for faster processing.

TABLE I. DATA SOURCES AND INFORMATION

Indicators	Reference Source	Frequency	Publisher	Scale	Information	Fields
Inflation	TABLE8_1 in the SEKI Report sector real.	Monthly	BI	Percentage	Monthly observations	Publish date, inflation_yoy, Inflation_mtm
Money Supply	TABLE1_1 SEKI in the Report sector monetary	Monthly	BI	Trillion Rupiah	Monthly observations	Publish date, M2
Exchange Rate	*Jakarta Interbank Spot Dollar Rate* (JISDOR).	Daily	BI	Rupiah per US Dollar	The monthly value is calculated as the average of daily observations.	Publish date, JISDOR
Interest Rate	*Jakarta Interbank Offered Rate* (JIBOR).	Daily	BI	Percentage	The monthly value is calculated as the average of daily observations.	Publish date, JIBOR
Wages	Rata-rata upah tenaga kerja dalam satu bulan.	Semi-annual	BPS	Rupiah	Semi-annual observations	Publish date, wage
Foreign Reserve Asset	TABLE5_9 in the SEKI Report sector external.	Monthly	BI	Billion US Dollar	Monthly observations	Publish date, Foreign reserve
Balance of Payments	TABLE5_1 in the SEKI Report sector external.	Quarterly	BI	Billion US Dollar	Quarterly variables are transformed into monthly frequency.	Publish date, Balance of payments, export, import
Gross Domestic Product	TABEL7_1, TABEL7_2, and TABLE7_6 in the SEKI Report sector real.	Quarterly	BI	Trillion Rupiah	Quarterly variables are transformed into monthly frequency.	Publish date, GDP deflator, GDP Nominal
Seasonal event	Holiday calendars indicating cultural events.	Event-based	Holiday Calendar	"1" or "0"		Publish date, event tagging

Table I presents the data sources and characteristics of the macroeconomic and seasonal variables used in this study, including data frequency, measurement scale, and relevant information for machine learning–based inflation forecasting.

The dataset consists of heterogeneous temporal frequencies, including daily (exchange rate and interest rate), monthly (inflation, money supply, and foreign reserves), quarterly (balance of payments and GDP), and semi-annual (wages) observations. To ensure temporal consistency for machine learning modeling, daily variables are aggregated into monthly averages, while quarterly variables are proportionally transformed into monthly frequency through time-series resampling techniques.

In addition, each variable is supported by structured fields such as publication date and relevant economic measures for easier tracking and integration. A temporal key is maintained with a variable "month", which is generated from the publish date to ensure that all indicators can be aggregated as one

analytical dataset. Seasonal effects are represented using dummy variables (1 and 0) for calendar-related events that may affect inflation patterns. A high-level overview of this well-structured dataset serves as a brief foundation for data preparation and modeling.

C. Data Preparation

This phase involves preparing the dataset through data cleaning, frequency harmonization, feature engineering, data integration, and variable transformation for machine learning analysis. These processes are necessary to ensure temporal consistency, improve data quality, and enhance model performance.

Several anomalous observations, particularly zero values in the interest rate series, are identified as outliers and removed prior to analysis. Since the dataset consists of heterogeneous temporal frequencies, including daily, monthly, quarterly, and semi-annual observations, frequency harmonization is performed using time-series resampling techniques. All variables are transformed into a monthly frequency to maintain temporal alignment across predictors and support consistent forecasting analysis.

TABLE II. FREQUENCY HARMONIZATION TECHNIQUES

Variable	Conversion	Resampling Technique
Exchange Rate	Daily to Monthly	Monthly average
Interest Rate	Daily to Monthly	Monthly average
Wages	Semi-annual to Monthly	Interpolation
Balance of Payments	Quarterly to Monthly	Proportionally distributed across three months
GDP	Quarterly to Monthly	Proportionally distributed across three months

Table II presents the frequency harmonization process using time-series resampling techniques to convert variables into a consistent monthly frequency.

The wage variable requires particular attention because it is originally available at a semi-annual frequency and subsequently transformed into monthly observations using interpolation. This approach preserves temporal consistency with other macroeconomic indicators and maintains dataset completeness for machine learning modeling. However, the interpolation process may introduce smoother transitions than actual short-term wage fluctuations, potentially reducing volatility within the series and affecting the sensitivity of the models to labor-related inflation dynamics.

After harmonization, all macroeconomic and seasonal variables are integrated into a unified dataset using the month variable as the temporal key. Feature engineering is then performed to generate several derived variables relevant to inflation dynamics, including Broad Money (BM), Velocity of Money (VM), Exchange Rate Pass-Through (ERPT), and Import Cover Ratio (ICR). At this stage, all derived variables are standardized to maintain scale consistency.

1) Broad Money (BM): Measures the level of liquidity in the economy by comparing money supply to annual nominal GDP. An excessively high liquidity ratio reflects money supply expansion beyond real activity and may generate inflationary pressure when not supported by productive capacity growth [38]. BM is calculated as [38]:

$$BM = \frac{Money\ supply\ (M2)}{Annual\ Nominal\ GDP}\ x\ 100 \quad (13)$$

2) Velocity of Money (VM): It is a measure of how quickly money is circulating within an economy over a period of time, and it is derived from Fisher's quantity theory of money. When the velocity of money increases, inflation tends to increase, and decreases may even dampen inflation when the money supply expands [3]. Thus, inflation analysis must consider not only money supply but also the velocity of money circulation [3].

$$VM = \frac{Annual\ Nominal\ GDP}{Money\ supply\ (M2)} \quad (14)$$

3) Exchange Rate Pass Through (ERPT): A metric of the extent to which the exchange rate changes are passed through to the domestic prices. A value of one indicates full transmission, and a value of less than one means incomplete passing through. ERPT consists of the ratio between changes in domestic prices and changes in the exchange rate expressed as elasticity.

$$ERPT = \frac{\%\ \triangle P(Domestic\ price\ changes)}{\%\ \triangle ER\ (Changes\ in\ exchange\ rates)} \quad (15)$$

4) Import Cover Ratio (ICR): Measures the ability of foreign exchange reserves to finance a country's imports without additional inflows. A high ICR supports exchange rate stability and helps contain inflation by limiting increases in import prices, whereas a low ICR constrains reserve adequacy, increases depreciation risk, raises import costs, and amplifies inflationary pressures [39].

$$ICR = \frac{Foreign\ Reserve\ Asset}{Monthly\ import\ value} \quad (16)$$

In addition, temporal variables such as month, quarter, and year are constructed to capture potential time-series patterns and seasonal effects not fully represented by macroeconomic indicators alone.

The dataset is subsequently cleaned to ensure consistency and modeling compatibility. Missing values are handled selectively depending on the characteristics of each variable, while irrelevant attributes used during preprocessing are removed to improve modeling efficiency. Prior to modeling, the dataset is organized chronologically to preserve temporal order and avoid potential data leakage in the forecasting process.

D. Modeling

The modeling phase aims to develop an interpretable machine learning framework for inflation forecasting by implementing Random Forest, XGBoost, and Artificial Neural Network (ANN) algorithms. These algorithms are selected because of their capability to capture nonlinear relationships and complex interactions among macroeconomic variables that influence inflation dynamics.

To systematically evaluate forecasting performance, the models are assessed under three experimental scenarios: 1)

baseline modeling using macroeconomic and seasonal variables, 2) an extended model incorporating time-series features through the inclusion of quarter variables, and 3) an enhanced model combining time-series features with hyperparameter optimization to improve predictive performance. This multi-scenario framework evaluates the contribution of feature augmentation and model optimization within a unified experimental setting.

The baseline scenario employs domestic macroeconomic indicators, including exchange rate (JISDOR), interest rate (JIBOR), money supply (M2), foreign reserves, balance of payments, GDP deflator, wages, Broad Money, Velocity of Money, ERPT, and Import Cover Ratio. Seasonal variables such as Ramadan, Eid al-Fitr, Christmas and New Year, and back-to-school periods are also included as predictors. The target variable is annual inflation (Inflation_yoy).

To preserve temporal consistency and avoid data leakage in time-series forecasting, the dataset is divided chronologically, where the earlier 80% of observations are used for training and the remaining 20% are reserved for out-of-sample testing. This chronological split ensures that future observations are not used during model training.

The second scenario extends the baseline configuration by incorporating the quarter variable as an additional predictor to capture potential temporal and seasonal patterns not fully represented by macroeconomic variables. The third scenario further extends the second model by implementing hyperparameter optimization using GridSearchCV to improve predictive accuracy and model stability.

Fig. 2 presents the implementation code of the baseline Random Forest model using the RandomForestRegressor function with default parameter settings.

```
# Selecting all column as predictors
X=df[['JISDOR', 'JIBOR', 'M2', 'Foreign Reserve', 'Balance of Payment',
      'GDP Deflator', 'Wage', 'Back To School', 'Christmas and New Year', 'Ramadhan',
      'Idul Fitri', 'Broad Money', 'ERPT', 'Velocity of Money', 'Impor Cover Ratio']]

# Specify 'Inflation_yoy' column as 'y'
y=df['Inflation_yoy']

# #Splitting the dataset into training and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train RandomForest model
model = rf.RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

#Making predictions on the test dataset
predictions = model.predict(X_test)
```

Fig. 2. Random Forest model in the baseline scenario.

Fig. 3 presents the Random Forest hyperparameter optimization using GridSearchCV with 5-fold cross-validation. Several parameters, including n_estimators, max_depth, min_samples_split, and min_samples_leaf, are systematically tuned to identify the optimal forecasting configuration. The best parameters obtained are n_estimators = 300, max_depth = None, min_samples_split = 2, and min_samples_leaf = 1.

Fig. 4 presents the implementation code of the baseline XGBoost model using the XGBRegressor function with default parameter settings.

```
from sklearn.model_selection import GridSearchCV
param_tuning = {
    'n_estimators': [100, 200, 300],
    'max_depth': [5, 10, 15, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
rf_model = rf.RandomForestRegressor(random_state=42)
grid_search = GridSearchCV(estimator = rf_model,
                           param_grid = param_tuning,
                           cv=5,
                           n_jobs = -1,
                           verbose = 1)
grid_search.fit(X_train, y_train)
print("Best parameters for RandomForestRegressor:", grid_search.best_params_)

# Fitting 5 folds for each of 108 candidates, totalling 540 fits
# Best parameters for RandomForestRegressor: {'max_depth': None,
#          'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
```

Fig. 3. Parameter optimization of Random Forest model.

```
# Selecting all column as predictors
X=df[['JISDOR', 'JIBOR', 'M2', 'Foreign Reserve', 'Balance of Payment',
      'GDP Deflator', 'Wage', 'Back To School', 'Christmas and New Year', 'Ramadhan',
      'Idul Fitri', 'Broad Money', 'ERPT', 'Velocity of Money', 'Impor Cover Ratio']]
# Specify 'Inflation_yoy' column as 'y'
y=df['Inflation_yoy']
# #Splitting the dataset into training and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train XGBoost model
model = xgb.XGBRegressor()
model.fit(X_train, y_train)
#Making predictions on the test dataset
predictions = model.predict(X_test)
```

Fig. 4. XGBoost model in the baseline scenario.

Fig. 5 presents the XGBoost hyperparameter optimization using GridSearchCV with 5-fold cross-validation. Several parameters, including learning_rate, max_depth, min_child_weight, subsample, colsample_bytree, and n_estimators, are systematically tuned to identify the optimal forecasting configuration. The best parameters obtained are learning_rate = 0.1, max_depth = 5, min_child_weight = 5, n_estimators = 200, subsample = 0.7, and colsample_bytree = 0.5.

```
# Hyperparameter tuning
from xgboost import XGBRegressor
from sklearn.model_selection import GridSearchCV

param_tuning = {
    'learning_rate': [0.01, 0.1],
    'max_depth': [3, 5, 7, 10],
    'min_child_weight': [1, 3, 5],
    'subsample': [0.5, 0.7],
    'colsample_bytree': [0.5, 0.7],
    'n_estimators' : [100, 200, 500],
    'objective': ['reg:squarederror']
}
xgb_model = XGBRegressor()
grid_search = GridSearchCV(estimator = xgb_model,
                           param_grid = param_tuning,
                           cv=5,
                           n_jobs = -1,
                           verbose = 1)
grid_search.fit(X_train, y_train)
print("Best parameters:", grid_search.best_params_)

# Fitting 5 folds for each of 288 candidates, totalling 1440 fits
# Best parameters: {'colsample_bytree': 0.5, 'learning_rate': 0.1, 'max_depth': 5,
#   'min_child_weight': 5, 'n_estimators': 200, 'objective': 'reg:squarederror', 'subsample': 0.7}
```

Fig. 5. Parameter optimization of XGBoost model.

The ANN model is implemented using the MLPRegressor function. Prior to training, feature normalization is performed using the Min–Max Scaling to transform all variables into the range [0,1]. This step is important for improving convergence stability and preventing variables with larger scales from dominating the learning process. Fig. 6 presents the implementation code of the ANN model in the baseline scenario.

```
# Selecting all column as predictors
X=df[['JISDOR', 'JIBOR', 'M2', 'Foreign Reserve', 'Balance of Payment',
      'GDP Deflator', 'Wage', 'Back To School', 'Christmas and New Year', 'Ramadhan',
      'Idul Fitri', 'Broad Money', 'ERPT', 'Velocity of Money', 'Impor Cover Ratio']]
# Specify 'Inflation_yoy' column as 'y'
y=df['Inflation_yoy']
# #Splitting the dataset into training and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Normalize dataset
scaler = MinMaxScaler()
# Fit and transfrom
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
# Train model with default parameter
model = MLPRegressor(random_state=42)
model.fit(X_train, y_train)
# Making predictions on the test dataset
predictions = model.predict(X_test)
```

Fig. 6. ANN model in the baseline scenario.

Fig. 7 presents the ANN hyperparameter optimization using GridSearchCV with 3-fold cross-validation. Several parameters, including hidden_layer_sizes, activation, solver, and max_iter, are systematically tuned to identify the optimal forecasting configuration. The best parameters obtained are hidden_layer_sizes = (256, 128, 64), activation = tanh, solver = adam, and max_iter = 1000, indicating that deeper network architectures and sufficient training iterations improve the predictive performance of the ANN model.

```
from sklearn.model_selection import GridSearchCV

# Define the parameter grid to search
param_grid = {
    'hidden_layer_sizes': [(100,), (50, 50), (256, 128, 64)],
    'activation': ['relu', 'tanh'],
    'solver': ['adam'],
    'max_iter': [500, 1000, 1500]
}
# Initialize MLPRegressor
mlp = MLPRegressor(random_state=42)
# Initialize GridSearchCV
grid_search = GridSearchCV(mlp, param_grid, cv=3,
                           scoring='neg_mean_squared_error', verbose=2, n_jobs=-1)
# Fit GridSearchCV to the training data
grid_search.fit(X_train, y_train)
# Print the best parameters and best score
print("Best parameters found: ", grid_search.best_params_)

# Fitting 3 folds for each of 18 candidates, totalling 54 fits
# Best parameters found:  {'activation': 'tanh', 'hidden_layer_sizes': (256, 128, 64),
#                'max_iter': 1000, 'solver': 'adam'}
```

Fig. 7. Parameter optimization of ANN model.

E. Evaluation

The evaluation phase assesses the performance and reliability of the developed machine learning forecasting models. Model performance is evaluated on out-of-sample testing data using standard forecasting accuracy metrics, including RMSE, MAE, and R^2.

Root Mean Squared Error (RMSE) is a widely used metric for evaluating predictive accuracy in statistical and machine learning models. It is defined as the square root of the average squared difference between actual and predicted values. A lower RMSE indicates better model performance, as larger errors are penalized more heavily.

Mean Absolute Error (MAE) measures the average magnitude of errors between actual and predicted values. A lower MAE indicates better predictive performance. Since it is expressed in the same unit as the target variable, MAE is easy to interpret and suitable for model comparison.

Coefficient of determination (R^2) is a statistical metric used to evaluate regression model performance by measuring the proportion of variance in the dependent variable explained by the independent variables. The value of R^2 ranges from 0 to 1, where values closer to 1 indicate better explanatory power and model performance.

To further validate forecasting performance differences, statistical significance testing is conducted using the Diebold–Mariano (DM) test. The test evaluates whether forecast error differences between Random Forest, XGBoost, and ANN models are statistically significant.

In addition, XAI techniques are employed through SHAP (Shapley Additive Explanations) value analysis to quantify the contribution of each predictor variable, thereby ensuring that the forecasting framework is not only predictive but also economically interpretable.

F. Deployment

The Deployment phase represents the final stage in which the selected model is prepared for practical use. In this study, deployment is conducted within an analytical and decision-support context by utilizing the best-performing model for inflation forecasting and forward-looking monetary analysis. The deployment process includes presenting forecasting and interpretability outputs to support transparent inflation analysis. SHAP analysis is applied to explain the contribution of each predictor variable to forecasting outcomes. In addition, the proposed framework supports periodic updates using new data, enabling the model to remain adaptive to evolving macroeconomic conditions and inflation dynamics.

IV. Results and Discussion

This section presents the performance evaluation results of the machine learning models across the three experimental scenarios, followed by interpretability analysis using SHAP and feature importance. SHAP analysis is applied to improve interpretability, while feature importance is used to support and validate the SHAP findings. The analysis evaluates forecasting performance, examines the effect of temporal feature augmentation and hyperparameter optimization, and identifies the macroeconomic variables that most strongly influence inflation predictions.

The regression performance of each model is evaluated using RMSE, MAE, and R^2 metrics to measure prediction accuracy, forecasting error, and explanatory power. Lower RMSE and MAE values indicate better forecasting performance, while higher R^2 values indicate stronger explanatory capability. In addition, Diebold–Mariano (DM) statistical testing is conducted to examine whether forecasting performance differences between models are statistically significant.

TABLE III. PERFORMANCE EVALUATION SUMMARY

Model	Scenario	RMSE	MAE	R^2
Random Forest	Baseline (macroeconomic indicators and seasonal variables)	0.5145	0.3619	0.9486
	Baseline Configuration with Time-Series Features	0.5287	0.3758	0.9457
	Baseline Configuration with Time-Series Features and Hyperparameter Tuning	0.5005	0.3469	0.9513
XGBoost	Baseline (macroeconomic indicators and seasonal variables)	0.6541	0.3664	0.9169
	Baseline Configuration with Time-Series Features	0.6245	0.3603	0.9242
	Baseline Configuration with Time-Series Features and Hyperparameter Tuning	**0.4987**	**0.3339**	**0.9517**
Artificial Neural Network	Baseline (macroeconomic indicators and seasonal variables)	1.7523	1.3153	0.4035
	Baseline Configuration with Time-Series Features	1.6543	1.2727	0.4683
	Baseline Configuration with Time-Series Features and Hyperparameter Tuning	1.0065	0.6503	0.8031

Table III presents the comparative performance of Random Forest, XGBoost, and ANN models across all scenarios using RMSE, MAE, and R^2 metrics.

The results consistently indicate that ensemble-based models outperform ANN across all experimental scenarios. In the baseline configuration, Random Forest achieves the strongest predictive performance, effectively capturing nonlinear relationships among macroeconomic and seasonal variables. XGBoost also performs relatively well, while ANN shows substantially weaker forecasting performance.

The weaker ANN performance in the baseline scenario may be associated with the sensitivity of neural networks to architecture configuration, parameter initialization, training iterations, and dataset size. Compared with ensemble-based models, ANN generally requires more extensive optimization to effectively learn nonlinear macroeconomic relationships. In contrast, tree-based ensemble methods are typically more robust for structured tabular economic data.

The inclusion of temporal variables in the second scenario produces mixed results. Random Forest shows a slight decline in performance, while XGBoost demonstrates moderate improvement after incorporating the quarter variable. ANN also shows limited improvement but remains below ensemble-based methods. These findings indicate that temporal feature augmentation is model-dependent rather than universally beneficial.

The best overall performance is achieved in the third scenario after hyperparameter optimization. XGBoost becomes the best-performing model with RMSE = 0.4987, MAE = 0.3339, and R^2 = 0.9517, closely followed by Random Forest with RMSE = 0.5005 and R^2 = 0.9513. ANN also improves substantially after optimization, confirming that neural network performance is highly dependent on architecture selection and parameter tuning. Overall, the findings indicate that hyperparameter optimization contributes more significantly to forecasting improvement than temporal feature augmentation.

To further validate forecasting performance differences between models, Diebold–Mariano (DM) testing is conducted using prediction errors from the testing dataset. The results indicate that the forecasting performance difference between XGBoost and Random Forest is not statistically significant (p-value > 0.05), suggesting that both ensemble-based models achieve relatively comparable predictive performance. However, both XGBoost and Random Forest significantly outperform ANN (p-value < 0.05), confirming the superior forecasting capability of ensemble-based methods for structured macroeconomic inflation data.

TABLE IV. DIEBOLD–MARIANO (DM) STATISTICAL TESTING

Comparison	DM Statistic	p-value	Result
XGBoost vs. RF	-1.7644	0.0851	Not Significant
XGBoost vs. ANN	-2.7265	0.0094	Significant
RF vs. ANN	-2.6407	0.0117	Significant

Table IV presents the results of the Diebold–Mariano (DM) statistical testing used to evaluate the statistical significance of forecasting performance differences between XGBoost, Random Forest, and ANN models.

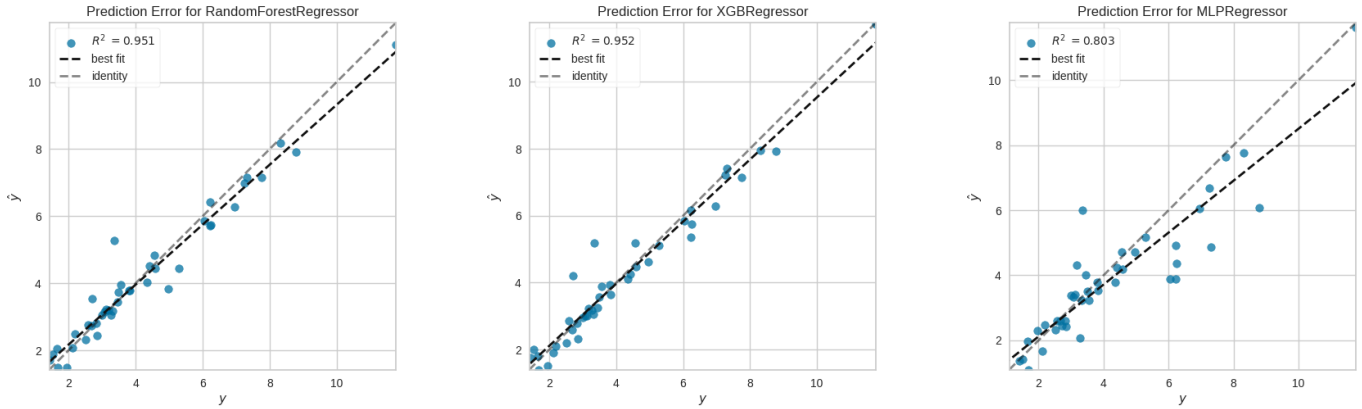


Fig. 8. Comparison of prediction error plots for Random Forest, XGBoost, and MLP (ANN) models in the third scenario.

Fig. 8 presents the prediction error plots for the three models in the third scenario. Random Forest and XGBoost show strong alignment between actual and predicted values, with most observations distributed near the identity line, indicating high predictive precision and limited systematic error. In contrast, ANN exhibits wider dispersion around the identity line, reflecting weaker predictive stability and larger forecasting errors.

To improve interpretability, SHAP analysis is applied only to the best-performing XGBoost model in the third scenario to provide transparent and policy-relevant insights into the contribution of macroeconomic variables to inflation forecasting. The analysis explains both the contribution and directional impact of each predictor variable on inflation forecasts, thereby helping to identify the key economic factors driving inflation.

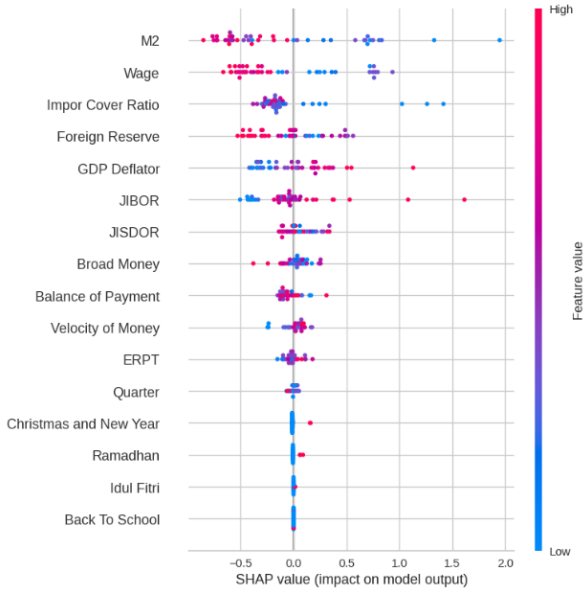


Fig. 9. SHAP summary for XGBoost model (third scenario).

Fig. 9 presents SHAP analysis for the XGBoost model (third scenario), showing that M2, Wage, and Import Cover Ratio are the main drivers, while other variables have moderate to low impact.

The dominance of M2 indicates the importance of monetary liquidity in influencing inflation dynamics, while Wage reflects labor-cost and purchasing-power effects. In addition, the relatively strong contribution of the Import Cover Ratio suggests that external-sector resilience and reserve adequacy also influence inflation stability through exchange-rate and import-price channels. Seasonal and temporal variables contribute relatively less, indicating that structural macroeconomic conditions play a more important role than calendar-based effects.

To support the SHAP results, feature importance analysis is subsequently performed for the best-performing XGBoost model. This additional analysis is intended to validate the consistency of the SHAP findings by identifying the variables that contribute most significantly to the forecasting process from the model perspective. The comparison between SHAP and feature importance provides complementary interpretability insights and strengthens the reliability of the identified macroeconomic determinants of inflation.

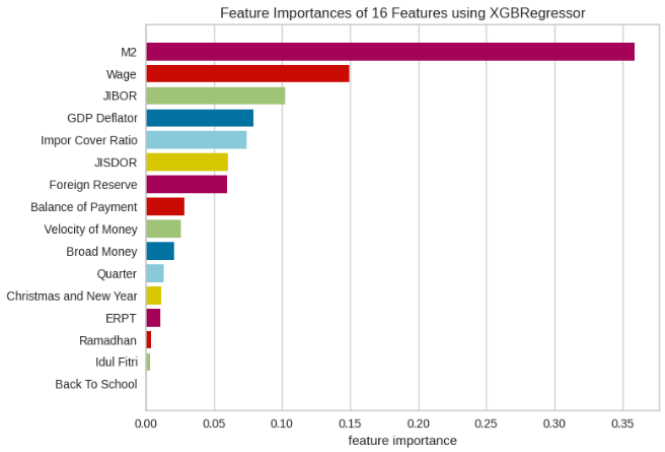


Fig. 10. Feature importance of XGBoost model (third scenario).

Fig. 10 presents the feature importance analysis for the best-performing XGBoost model in the third scenario. The results show that M2 is the most influential variable, indicating that monetary liquidity plays a dominant role in inflation forecasting. Wage emerges as the second most important variable, suggesting that labor-cost and purchasing-power effects also contribute significantly to inflation dynamics. Other variables, including JIBOR, GDP Deflator, and Import Cover Ratio, exhibit moderate importance, reflecting the influence of interest rate transmission, general price-level conditions, and external-sector resilience on inflation behavior.

Compared with the SHAP analysis of the best-performing XGBoost model, the feature importance results show a consistent pattern. Both analysis identify M2 and Wage as the dominant contributors to inflation forecasting, confirming the strong influence of monetary liquidity and labor-related factors on inflation dynamics in Indonesia. In addition, Import Cover Ratio also appears as an influential variable, highlighting the role of external-sector stability and reserve adequacy in influencing inflation through exchange-rate and import-price channels.

However, SHAP provides deeper interpretability than feature importance because it explains not only the magnitude but also the direction of each variable's contribution to inflation predictions. While feature importance identifies globally influential variables, SHAP further explains how variables such as M2 and Wage positively or negatively affect predictions across observations. The consistency between both analyses strengthens the reliability of the interpretability findings and confirms that inflation dynamics in Indonesia are primarily driven by structural macroeconomic conditions rather than seasonal effects.

V. Conclusion

This study proposes an interpretable machine learning framework for inflation forecasting in Indonesia using domestic macroeconomic and seasonal variables. The framework systematically evaluates Random Forest, XGBoost, and ANN

models under three experimental scenarios: baseline modeling, temporal feature augmentation, and hyperparameter optimization.

The results consistently show that ensemble-based methods outperform ANN across all scenarios, with XGBoost achieving the best overall performance after hyperparameter optimization (RMSE = 0.4987, MAE = 0.3339, R^2 = 0.9517), closely followed by Random Forest (RMSE = 0.5005, R^2 = 0.9513). Diebold–Mariano (DM) testing further indicates that the forecasting performance difference between XGBoost and Random Forest is not statistically significant, whereas both ensemble-based models significantly outperform ANN. These findings confirm the effectiveness of ensemble-based machine learning models for inflation forecasting.

To improve interpretability, SHAP analysis is applied to the best-performing XGBoost model. The results identify M2, Wage, and Import Cover Ratio as the dominant contributors to inflation forecasting, indicating that monetary liquidity, labor-related factors, and external-sector resilience play important roles in explaining inflation dynamics in Indonesia.

Overall, this study contributes through a systematic forecasting framework that integrates predictive accuracy, temporal feature evaluation, hyperparameter optimization, statistical validation, and explainable artificial intelligence within a unified experimental setting. From a practical perspective, the proposed framework may support more transparent and forward-looking monetary policy analysis in Indonesia and other emerging economies with similar economic structures.

Nevertheless, this study has several limitations. It focuses primarily on domestic macroeconomic indicators and does not yet incorporate broader external variables. In addition, the harmonization of lower-frequency variables into monthly frequency may simplify some short-term dynamics. Future studies may incorporate developed-country interest rates and global commodity prices to better capture external monetary influences and global shocks affecting inflation dynamics.

REFERENCES

[1] H. Z. Mustofa and Moh. Faizin, "Effect of Macroeconomic Factors on Economic Growth in Indonesia," Journal of Developing Economies, vol. 10, no. 1, pp. 48–72, Jun. 2025, doi: 10.20473/jde.v10i1.60620.

[2] A. Bilalli, M. Sadiku, and L. Sadiku, "THE IMPACT OF INFLATION ON FINANCIAL SECTOR PERFORMANCE: EVIDENCE FROM OECD COUNTRIES," ECONOMICS - Innovative and Economics Research Journal, vol. 12, no. 2, pp. 263–276, Aug. 2024, doi: 10.2478/eoik-2024-0028.

[3] D. Hermawan, D. Lie, A. Sasongko, and R. I. Yusan, "Money velocity, digital currency and inflation dynamics in Indonesia," 2024, Taylor and Francis Ltd. doi: 10.1080/00074918.2024.2398347.

[4] S. Isnowati, F. Sugiyanto, A. S. Kurnia, and E. Tjahjaningsih, "The Effect Labor Wage and Exchange Rate on Inflation," Montenegrin Journal of Economics, vol. 19, no. 1, pp. 117–126, 2023, doi: 10.14254/1800-5845/2023.19-1.10.

[5] H. Harmanta, M. B. Bathaluddin, and J. Waluyo, "INFLATION TARGETING UNDER IMPERFECT CREDIBILITY: LESSONS FROM INDONESIAN EXPERIENCE," Buletin Ekonomi Moneter dan Perbankan, vol. 13, no. 3, pp. 271–306, May 2011, doi: 10.21098/bemp.v13i3.394.

[6] V. Fibriyani and N. Chamidah, "Prediction of Inflation Rate in Indonesia Using Local Polynomial Estimator for Time Series Data," in Journal of Physics: Conference Series, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1776/1/012065.

[7] A. A. Naghi, E. O'Neill, and M. Danielova Zaharieva, "The benefits of forecasting inflation with machine learning: New evidence," Journal of Applied Econometrics, vol. 39, no. 7, pp. 1321–1331, Nov. 2024, doi: 10.1002/jae.3088.

[8] Y.-S. Li, P.-F. Pai, and Y.-L. Lin, "Forecasting inflation rates be extreme gradient boosting with the genetic algorithm," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 3, pp. 2211–2220, Mar. 2023, doi: 10.1007/s12652-022-04479-4.

[9] K. Amarasinghe, K. T. Rodolfa, H. Lamba, and R. Ghani, "Explainable machine learning for public policy: Use cases, gaps, and research directions," Data Policy, vol. 5, Feb. 2023, doi: 10.1017/dap.2023.2.

[10] W. Badrawani, "An Interpretable Machine Learning Approach in Predicting Inflation Using Payments System Data: A Case Study of Indonesia," Jun. 2025, doi: 10.48550/arXiv.2506.10369.

[11] S. S. Sharma, "Which variables predict Indonesia's inflation?," Buletin Ekonomi Moneter dan Perbankan/Monetary and banking economics bulletin, vol. 22, no. 1, pp. 87–101, 2019, doi: 10.21098/bemp.v22i1.1038.

[12] I. A. Moosa, K. Al-Saad, and I. N. Khatatbeh, "The Quantity Theory of Money, Quantitative Easing and the Missing Inflation Phenomenon," Journal of Central Banking Theory and Practice, vol. 13, no. 2, pp. 71–88, May 2024, doi: 10.2478/jcbtp-2024-0013.

[13] T. Stylianou, R. Nasir, and M. Waqas, "The relationship between money supply and inflation in Pakistan," PLoS One, vol. 19, no. 3 March, Mar. 2024, doi: 10.1371/journal.pone.0301257.

[14] U. Shahzad, B. Orsi, and G. D. Sharma, "Managing inflation expectations and the efficiency of monetary policy responses to energy crises," Energy Econ., vol. 133, May 2024, doi: 10.1016/j.eneco.2024.107474.

[15] V. B. Nguyen, "Effects of fiscal deficit and money M2 supply on inflation: Evidence from selected economies of Asia," Journal of Economics, Finance and Administrative Science, vol. 20, no. 38, pp. 49–53, Jun. 2015, doi: 10.1016/j.jefas.2015.01.002.

[16] D. H. Vo and N. T. Vu, "Exchange rate pass-through for European Union countries," PLoS One, vol. 19, no. 8 August, Aug. 2024, doi: 10.1371/journal.pone.0309527.

[17] J. Durcova, "The impact of interest rate transmission channel on the prices development in the eurozone countries," Montenegrin Journal of Economics, vol. 17, no. 2, pp. 23–35, 2021, doi: 10.14254/1800-5845/2021.17-2.2.

[18] A. A. Iddrisu and I. P. Alagidede, "Revisiting interest rate and lending channels of monetary policy transmission in the light of theoretical prescriptions," Central Bank Review, vol. 20, no. 4, pp. 183–192, Dec. 2020, doi: 10.1016/j.cbrev.2020.09.002.

[19] B. D. Krušković, "Econometric VAR Analysis of the Effect of the Foreign Exchange Reserves on Macroeconomic Variables in Emerging Countries: The Case of BRIC Countries," Journal of Central Banking Theory and Practice, vol. 12, no. 3, pp. 87–109, Sep. 2023, doi: 10.2478/jcbtp-2023-0026.

[20] E. E. Sumiyati, "Macroeconomic Factors and Current Account Deficit in Indonesia," J. Econ. Bus., vol. 5, no. 4, Dec. 2022, doi: 10.31014/aior.1992.05.04.455.

[21] D. Peng, Y. Ye, and Q. Chen, "Impact of cross-border capital flows on foreign exchange market stability," Financ. Res. Lett., vol. 62, p. 105155, Apr. 2024, doi: 10.1016/j.frl.2024.105155.

[22] R. Kollmann, "Effects of Covid-19 on Euro area GDP and inflation: demand vs. supply disturbances," International Economics and Economic Policy, vol. 18, no. 3, pp. 475–492, Jul. 2021, doi: 10.1007/s10368-021-00516-3.

[23] L. Jin, Z. Liu, and L. Li, "Prediction and identification of nonlinear dynamical systems using machine learning approaches," J. Ind. Inf. Integr., vol. 35, p. 100503, Oct. 2023, doi: 10.1016/j.jii.2023.100503.

[24] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[25] Z. Farhadi, H. Bevrani, M. R. Feizi-Derakhshi, W. Kim, and M. F. Ijaz, "An Ensemble Framework to Improve the Accuracy of Prediction Using

Clustered Random-Forest and Shrinkage Methods," Applied Sciences (Switzerland), vol. 12, no. 20, Oct. 2022, doi: 10.3390/app122010608.

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[27] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, and O. Lyashevska, "Predictive analytics with gradient boosting in clinical medicine," Ann. Transl. Med., vol. 7, no. 7, pp. 152–152, Apr. 2019, doi: 10.21037/atm.2019.03.29.

[28] K. Sako, B. N. Mpinda, and P. C. Rodrigues, "Neural Networks for Financial Time Series Forecasting," Entropy, vol. 24, no. 5, May 2022, doi: 10.3390/e24050657.

[29] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, doi: https://doi.org/10.48550/arXiv.1705.07874.

[30] A. Sánchez, C. Vidal-Silva, G. Mancilla, M. Tupac-Yupanqui, and J. M. Rubio, "Sustainable e-Learning by Data Mining—Successful Results in a Chilean University," Sustainability (Switzerland), vol. 15, no. 2, Jan. 2023, doi: 10.3390/su15020895.

[31] M. F. Abd El-Aal, "Analysis Factors Affecting Egyptian Inflation Based on Machine Learning Algorithms," Data Science in Finance and Economics, vol. 3, no. 3, pp. 285–304, 2023, doi: 10.3934/DSFE.2023017.

[32] P. Sharma, A. Sharma, H. Manuj, A. Kambhampati, and A. Singh, "Machine Learning Applications for Predicting the Inflation Rate in India," in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2025, pp. 233–243. doi: 10.1007/978-981-96-5955-5_20.

[33] G. S. Araujo and W. P. Gaglianone, "Machine learning methods for inflation forecasting in Brazil: New contenders versus classical models," Latin American Journal of Central Banking, vol. 4, no. 2, p. 100087, Jun. 2023, doi: 10.1016/j.latcb.2023.100087.

[34] Y. Jouilil and M. Iaousse, "Comparing the Accuracy of Classical and Machine Learning Methods in Time Series Forecasting: A Case Study of USA Inflation," Statistics, Optimization & Information Computing, vol. 11, no. 4, pp. 1041–1050, Aug. 2023, doi: 10.19139/soic-2310-5070-1767.

[35] T. G. Baidoo and A. Obeng, "Navigating Inflation in Ghana: How Can Machine Learning Enhance Economic Stability and Growth Strategies," Oct. 2024, doi: 10.48550/arXiv.2410.05630.

[36] I. Seo, M. Kim, J. Wook Kim, and B. Jang, "Accurate total consumer price index forecasting with data augmentation, multivariate features, and sentiment analysis: A case study in Korea," PLoS One, vol. 20, no. 5, p. e0321530, May 2025, doi: 10.1371/journal.pone.0321530.

[37] M. Simionescu, "Machine Learning vs. Econometric Models to Forecast Inflation Rate in Romania? The Role of Sentiment Analysis," Mathematics, vol. 13, no. 1, p. 168, Jan. 2025, doi: 10.3390/math13010168.

[38] J. Nowicki, P. Ratajczak, and D. Szutowski, "Impact of Macroeconomic Factors on Financial Liquidity of Companies: A Moderation Analysis," Sustainability (Switzerland) , vol. 16, no. 11, Jun. 2024, doi: 10.3390/su16114483.

[39] S. K. Klutse, G. D. Kiss, and J. Sági, "Exchange market pressure in Sub-Saharan African countries – The role of imports and short-term external debt, 2002–2017," Regional Statistics, vol. 12, no. 4, pp. 92–111, 2022, doi: 10.15196/RS120404.

A Conceptual Model for Detecting Contactless Drug Distribution Based on Behavioural Analysis and Geospatial Visualisation

Medeu Kurmangali[1], Talgat Akimzhanov[2], Kanat Kazhibaev[3], Kulambayev Bakhytzhan[4]*, Moshkalov Altynbek[5]
Center for Military Strategic Research JSC, Astana, Kazakhstan[1]
Turan University, Almaty, Kazakhstan[2, 4]
Research & Development Centre Kazakhstan Engineering LLP, Astana, Kazakhstan[3]
Deputy Dean for Scientific Work and International Cooperation, PhD, Acting Associate Professor[5]

***Abstract*—This study proposes a multi-level system for detecting contactless drug distribution transactions. This system integrates behavioural pattern recognition as the primary detection channel, detection of night-time activity spikes as an enhancing module, and facial matching as an additional probabilistic evaluation layer. The system identifies a two-phase structure of covert transactions. Both the courier's placement of the product and the buyer's retrieval produce recognizable behavioural sequences at the same geographic location. Signal-to-noise ratio analysis identified a detection threshold at an SNR of approximately 17. It provides a quantitative foundation for camera placement planning. The behavioural pattern recognition pipeline integrates YOLO-Pose for skeleton estimation and employs classical machine learning models, including Random Forest and Gradient Boosting, for temporal action classification. The ST-GCN architecture is considered a future extension pending the availability of a larger annotated dataset.**

***Keywords*—Anomaly detection; flash detection; pose estimation; SNR; crime mapping**

I. Introduction

Contactless drug trafficking has become one of the fastest-growing modalities of retail drug distribution. According to the latest United Nations estimates, the global population of people who use drugs reached 292 million in 2022, and the ongoing digitalisation of drug markets has created new logistical patterns that challenge traditional law-enforcement tools [1]. Systematic reviews of dark-web cryptomarkets covering the period 2012–2023 show a consistent migration of retail drug distribution towards online ordering and anonymous offline delivery, supported by cryptocurrency-based payment and encrypted communications [2]. Cryptomarket participants explicitly design their workflow to minimise visible offline exposure. The qualitative analyses of vendor and buyer practices demonstrate that offline drop-offs and retrieval episodes remain the key points of contact at which law-enforcement detection is most feasible [3]. In the post-Soviet space, this distribution paradigm has been documented in qualitative harm-reduction studies that describe contactless handover (called "dead-drop") as the dominant channel for the supply of new psychoactive substances [4].

The operational cycle of contactless distribution involves two observable offline episodes. In the first, the courier leaves the package in a concealed location and photographs it to record the coordinates; in the second, the buyer visits the same location to pick up the goods. Both episodes result in a recognisable sequence of actions — approaching, stopping, crouching, manipulating objects at ground level, and leaving — repeated across individuals and locations. This behavioural repeatability is consistent with long-established findings in environmental criminology. It shows that a small number of "micro-places" concentrate a disproportionate share of criminal events and that routine offender behaviour clusters spatially in predictable patterns [5], [6], [7]. Recent systematic reviews of crime concentration confirm that the law of crime concentration at a place also extends to drug-related offences and can therefore be used to guide both detection and preventive patrol [8], [9].

This repetitive behavioural pattern represents the primary detection signal and is observable in CCTV footage regardless of the time of day. However, at night, the courier's use of a mobile phone camera flash to photograph the location of the stash introduces an additional photometric signal. Flashes create a short, spatially compact burst of brightness that can be distinguished from diffuse light sources such as car headlights. The temporal profile of LED-based flash emission has been experimentally characterised at high frame rates in fluid-dynamics imaging research [10].

Existing CCTV analytics systems focus on general anomaly detection [11], biometric identification [12], or crime prediction. A combination of behavioural, photometric, and geographic features specific to covert operations can improve analytical results. Advances in pose estimation and temporal action recognition [13], [14] enable the analysis of behavioural sequences from standard video streams, which paves the way for the development of a detection system adapted to this type of crime. Although numerous studies address general anomaly detection in surveillance videos, very few works analyse behavioural patterns associated with covert logistics operations such as contactless drug distribution. In addition, existing approaches rarely combine behavioural analysis with photometric signals such as smartphone flash events and geospatial recurrence of incidents.

This study proposes a three-layer framework. The framework includes behavioural pattern recognition as a core method operating in all lighting conditions, night flash detection

*Corresponding author

as a booster module, and face matching as a probabilistic evaluation feature. This study investigates two key components of the proposed framework: behavioural pattern recognition and flash detection. Behavioural analysis serves as the primary detection channel, while flash detection acts as an auxiliary photometric signal that increases confidence in low-light conditions.

This study makes four main contributions:

- A multimodal framework combining behavioural pattern recognition, photometric flash detection, and geospatial correlation for identifying contactless drug transactions.
- A statistically grounded flash detection model based on adaptive SNR thresholding and spatial blob validation.
- A semi-synthetic dataset generation approach for evaluating rare flash events under realistic surveillance noise conditions.
- An interpretable multimodal fusion model integrating behavioural and photometric signals for anomaly scoring.

II. Related Work

A. Contactless Drug Distribution and Place-Based Criminology

The criminological literature on micro-place concentration of crime provides a strong theoretical foundation for the problem addressed in this study. Sherman, Gartin, and Buerger showed that a small proportion of street addresses account for a disproportionately large share of calls for service [15]. A finding later formalised by Weisburd as the "law of crime concentration at place" and repeatedly replicated across cities and crime types [16], [9]. Braga, Papachristos, and Hureau extended this reasoning to serious violence, demonstrating the temporal stability of gun-violence hotspots over nearly three decades in Boston [6]. Kernel-density estimation and related hotspot mapping techniques have become standard instruments for translating such patterns into operational maps [17]. In the specific context of drug markets, contemporary research on cryptomarkets describes a "stretched" transaction structure in which the riskiest phases for participants are the packaging, drop-off, and retrieval steps that occur offline [18]. Subsequent work on dark-web trade trends [19] and qualitative studies of new psychoactive substances in Eurasia [4] confirm that anonymous offline drops are the principal vulnerability of contactless distribution to conventional policing.

B. Video Anomaly Detection

Anomaly detection in videos has continuously attracted attention in the computer-vision community. Sultani, Chen, and Shah introduced the UCF-Crime dataset containing 1 900 surveillance-camera videos with 13 anomaly categories and proposed a weakly supervised multiple-instance-learning (MIL) deep-learning framework that achieved an AUC of 0.75 on the full dataset [20]. Subsequent work by Tian et al. improved this baseline through a robust approach to learning the magnitude of temporal features [15]. Manju et al. combined a Mask Region-based Convolutional Neural Network (Mask R-CNN) with Long Short-Term Memory (LSTM) networks to provide early anomaly detection, reporting an accuracy of 93.6 % [16]. Khan et al. evaluated deep learning models for anomaly detection in traffic-surveillance videos using multi-stream architectures [17], while Li et al. proposed a context-related generative-adversarial-network approach to address the challenge of diverse scene conditions [18]. Phapale and Bhingarkar recently introduced a deep context-aware feature extractor that outperforms earlier MIL baselines on standard surveillance benchmarks [19], and Ullah et al. demonstrated that hybrid convolutional–transformer architectures further improve detection accuracy on UCF-Crime [21]. Broader reviews of the field [22-24] and a review of event detection in surveillance video [25] agree that, although individual components of anomaly detection, facial recognition, and crime mapping have been extensively studied. They have not been combined into a framework specifically designed to analyse the behavioural patterns of contactless drug trafficking.

C. Pose Estimation and Skeleton-Based Action Recognition

Skeleton-based action recognition has been advanced by the introduction of spatiotemporal graph-convolutional networks (ST-GCN) by Yan, Xiong, and Lin [13], which model joint relationships over time using graph convolutions. Later work refined this line of research through adaptive graph convolution [26], multi-stream spatiotemporal fusion [27], and graph-neural ordinary-differential-equation models [28]. Contemporary surveys of deep-learning-based pose estimation [29] show that high-resolution architectures such as HRNet [30] and real-time multi-person estimators such as OpenPose [31] form the current backbone of most downstream pose-based pipelines. The YOLO family of object detectors, including the recent YOLO11-Pose variants [32], enables real-time pose estimation, which is suitable for video-surveillance applications. Recent refinements of YOLOv8-Pose have reported improved keypoint accuracy under occlusion and partial visibility typical of CCTV scenes [33], [34], and dedicated YOLOv8-based pipelines have been successfully applied to suspicious-behaviour detection in public spaces [35]. For model evaluation, the NTU RGB+D and NTU RGB+D 120 datasets remain the de facto benchmarks for three-dimensional skeleton-based action recognition [36], [37], while the VIRAT corpus provides realistic multi-view outdoor surveillance video [38].

D. Biometric Identification and Face Recognition

Biometric identification plays a complementary role in the proposed system. Early multimodal biometric systems established fundamental design principles for combining heterogeneous identity cues [12]. Modern face-recognition research is dominated by angular-margin softmax losses, of which ArcFace is the canonical representative [39]. The Dyn-arcFace variant introduces a dynamic additive angular margin that adapts to the within-class structure of the training set [40] and is directly relevant to surveillance scenarios in which face quality is highly variable.

E. Positioning of this Work

The literature reviewed shows that while individual components—anomaly detection, pose-based action recognition, face recognition and crime mapping—have been extensively studied. Nevertheless, they have not been combined into a framework specifically designed to analyse the behavioural patterns of contactless drug trafficking. It appears

that prior work has not yet combined a statistically grounded flash-detection model with a skeleton-based behavioural classifier and a geospatial reporting layer. This study fills this gap by proposing an integrated approach that links incident detection, pose-based activity recognition, citizen reporting, and geospatial analysis under a single probabilistic decision rule.

III. Methodology

The proposed framework is organised into three functional modules that address complementary aspects of the detection problem: video-based recognition, citizen incident reporting, and geospatial analysis. Each module is implemented as stand-alone software and a toolkit, and registered as a utility model in the Republic of Kazakhstan. The modules share a common data layer via a centralised PostgreSQL database with the PostGIS extension for geographic queries.

A. System Architecture

The recognition module (registered as utility model KZ U 11660) processes media data to identify individuals potentially involved in drug trafficking. Its architecture consists of five components connected via HTTPS/SSL: 1) a media-data reception and verification server with subsystems for data input and format verification; 2) an intelligent processing server with four subsystems for motion-pattern analysis, results recording, model training, and authentication; 3) S3-compatible cloud object storage; 4) user client devices; and 5) analyst workstations. The module accepts photos and videos from citizens and uses computer-vision algorithms to detect suspicious behavioural patterns such as crouching, digging, and stealthy movements. The software is written in Python, has a total size of 23 MB, and its minimum server requirements are 4 CPU cores, 8 GB of RAM, and 10–50 GB of SSD storage.

B. Mobile Reporting Application

The mobile application (utility model KZ U 11661) allows citizens to submit contactless reports of suspected illegal drug trafficking. It is implemented as a progressive web application (PWA), accessible via a URL without installation from app stores. The server-side component includes six subsystems: message reception, media verification, data storage, staff authorisation, an administrative panel, and event registration. Messages contain structured fields for incident type (online sale, stash, direct transfer), detection date and time, addresses with integrated OpenStreetMap/Leaflet geolocation, optional contact information for subsequent tracking, and photo or video attachments. The front-end uses Vue.js/Nuxt.js, the back-end runs on Python/Django with Celery and Redis, and PostgreSQL serves as the database management system. The application supports both anonymous and identified reports, with HTTPS encryption for all data transfers.

C. Digital Crime Map

The digital crime map (utility model KZ U 11658) provides geospatial visualisation and analysis of reported incidents. It operates in three modes: a public map integrated into the report form for precise identification of incident locations. An operator map displaying individual reports filtered by time period, district, and microdistrict, and a heat map aggregating incident density for strategic analysis. The map is built on the Leaflet.js library using OpenStreetMap tile data (ODbL licence) and deployed on a Linux/Nginx server platform. The back-end integrates with PostGIS for geographic queries and supports near-real-time synchronisation with the central database. Kernel-density-estimation heat maps of this type have been shown to be effective instruments for operational crime-pattern visualisation at the micro-place level [41], [42].

D. Integrated Architecture

Fig. 1 presents the overall architecture of the proposed multimodal artificial intelligence framework for the detection and analysis of contactless drug-related activities. The system is organized as a hierarchical multi-layer pipeline integrating heterogeneous data modalities, analytical modules, multimodal fusion mechanisms, and decision-support components into a unified operational framework. At the highest level, the architecture incorporates three primary categories of data sources: encrypted messaging platforms, financial transaction records, and CCTV video streams. These sources collectively provide complementary textual, transactional, behavioural, and spatial information associated with covert distribution activities. The integration of heterogeneous modalities enables the system to overcome the limitations of single-source surveillance approaches and supports the generation of more reliable risk assessments under complex urban conditions.

The second layer of the architecture consists of modality-specific analytical modules designed to process each data stream independently before multimodal integration. The Natural Language Processing (NLP) module employs LLM- and BERT-based analysis techniques to perform message extraction, entity recognition, intent classification, and textual risk estimation from communication data. Simultaneously, the Financial Anomaly Detection module analyses transaction profiles and abnormal monetary patterns associated with suspicious operational behaviour. The Geospatial Analysis Module performs spatial clustering, hotspot detection, and movement-pattern analysis using machine-learning-based geospatial modelling techniques such as DBSCAN clustering and kernel density estimation (KDE). Each analytical branch generates modality-specific feature representations and intermediate risk scores, which are subsequently forwarded to the fusion stage. This modular architecture ensures scalability and allows each analytical component to be independently optimized or extended in future system versions.

The third layer represents the Multimodal Data Fusion Layer, which serves as the central integration component of the proposed framework. In this stage, modality-specific feature vectors $f_1, f_2, \ldots, f_n$ are normalized and aligned prior to fusion. The system supports both weighted and attention-based fusion strategies, enabling adaptive prioritization of heterogeneous evidence sources depending on operational conditions. The fusion process produces a unified risk score $R_{fusion} \in [0,1]$, which reflects the aggregated probability of suspicious activity. The integrated representation combines temporal dynamics, behavioral anomaly scores, spatial recurrence patterns, flash-detection confidence, textual risk indicators, and financial anomalies into a single probabilistic representation. Additionally, the architecture incorporates a feedback loop for model updating and continuous learning, enabling iterative

improvement of detection performance as new operational data become available.

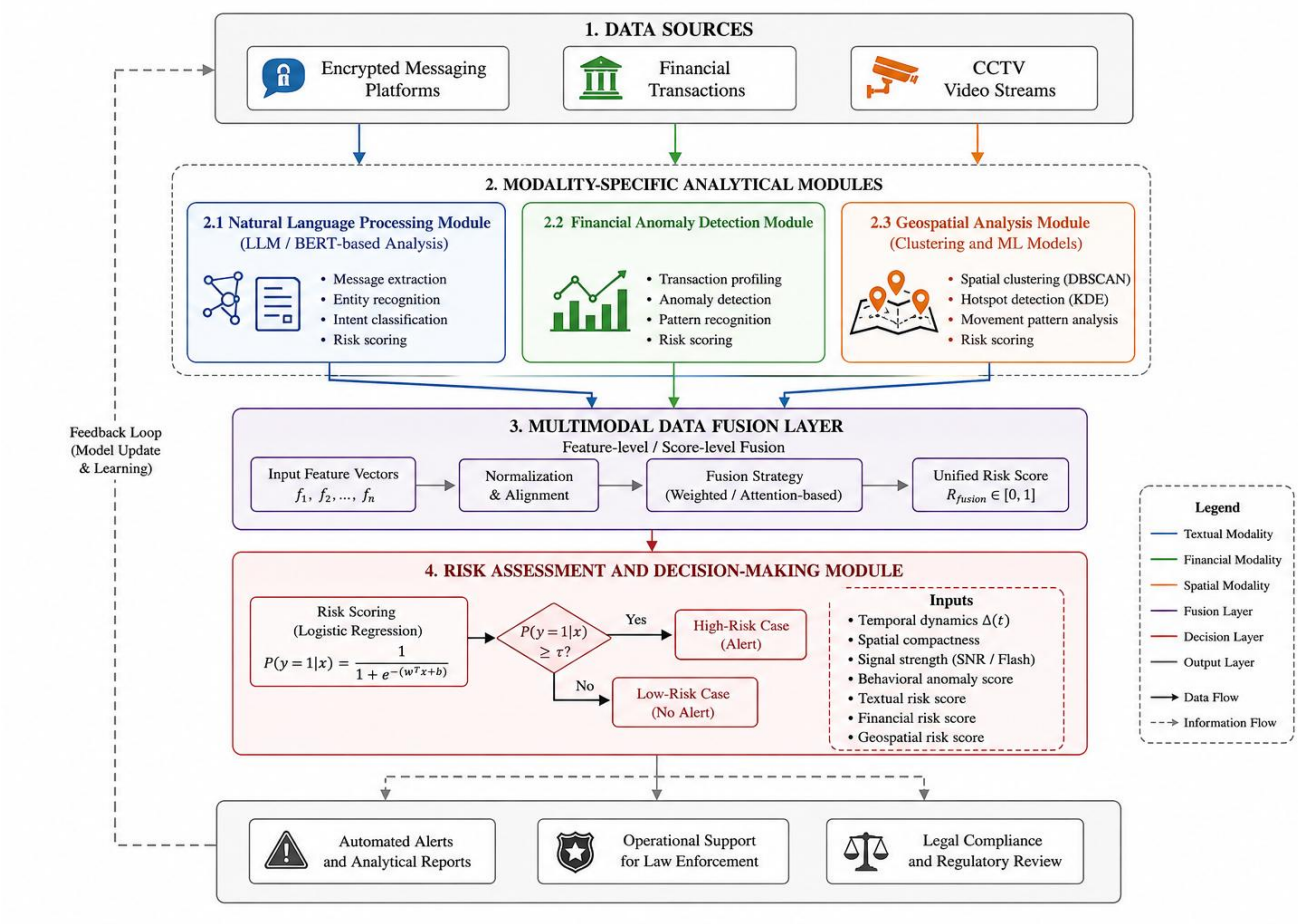


Fig. 1. Architecture of an integrated artificial intelligence system for contactless detection of drug-related crimes.

The final stage of the framework is the Risk Assessment and Decision-Making Module, which transforms fused multimodal evidence into operational decisions. A logistic-regression-based probabilistic scoring mechanism evaluates whether the estimated risk probability $P(y = 1 \mid x)$ exceeds a predefined threshold τ. Cases exceeding the threshold are classified as high-risk events and generate alerts for further investigation, while low-risk cases are filtered to reduce false-positive reporting. The decision layer integrates multiple inputs, including temporal intensity dynamics $\Delta(t)$, spatial compactness, signal-to-noise ratio (SNR), behavioral anomaly scores, and modality-specific risk estimates. The resulting outputs include automated alerts and analytical reports, operational support tools for law-enforcement agencies, and legal-compliance review mechanisms. Consequently, Fig. 1 demonstrates how the proposed framework combines multimodal artificial intelligence, geospatial analytics, behavioural modelling, and probabilistic decision-making into a comprehensive surveillance-oriented analytical system for detecting covert drug-distribution activities.

Three parallel AI modules process these data. Additional modules such as NLP and financial analysis are part of future system extensions and are not evaluated in this study. A multimodal fusion layer combines the outputs of these modules, and a risk-assessment module generates alerts, recommends law-enforcement actions, and conducts compliance checks.

E. Flash Detection Model

To formalise the detection problem, the observed video sequence is modelled as a spatio-temporal signal composed of background, signal, and noise components:

$$I(x, y, t) = B(x, y) + S(x, y, t) + N(x, y, t) \quad (1)$$

where, *B(x,y)* denotes the static background, *S(x,y,t)* represents the flash signal, and *N(x,y,t)* is an additive noise process. The noise is assumed to follow a Gaussian distribution:

$$N(x, y, t) \sim \mathcal{N}(0, \sigma^2)$$

The flash-detection problem can be formulated as a binary hypothesis-testing task under the classical Neyman–Pearson framework [40], [41]:

$$H_0: S(x, y, t) = 0, H_1: S(x, y, t) \neq 0$$

Under the Gaussian noise assumption, the inter-frame intensity variation *Δ(t)* follows:

$$\Delta(t) \sim \begin{cases} \mathcal{N}(0, \sigma^2), & H_0 \\ \mathcal{N}(\mu_s, \sigma^2), & H_1 \end{cases}$$

F. Behavioural Recognition Pipeline

Behavioural pattern recognition constitutes the core detection layer of the proposed system and operates independently of lighting conditions. Spatio-temporal graph convolutional networks (ST-GCN) represent the state-of-the-art approach for skeleton-based action recognition. This study focuses on classical machine learning models due to the limited size of the labelled dataset. The processing pipeline, shown in Fig. 2, processes raw video footage from CCTV cameras through six sequential layers to identify a characteristic sequence of actions associated with dead drop transactions.

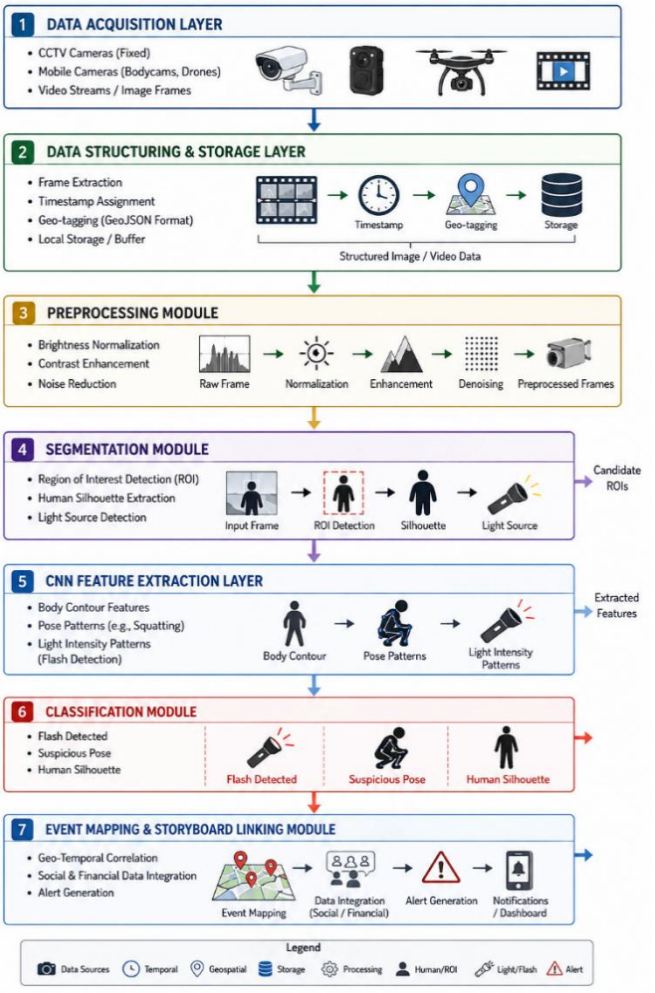


Fig. 2. Architecture of the deep learning visual layer for video analysis.

Notably, this sequence is repeated twice in each transaction cycle: first when the courier places the dead drop, and again when the buyer arrives to pick it up. The repetition of structurally similar behaviour in the same geographic location by different individuals strengthens the detection signal and allows the system to link related events over time.

The data-collection layer receives streams from fixed CCTV cameras and mobile sources. Training data are collected from three sources: open-source video footage, completed criminal case files, and scenarios involving actors. The scenarios are divided into target actions (placing a hiding place, taking a photo, examining the surroundings, picking up a hidden object) and hard-negative actions (tying shoelaces, picking up dropped keys, cleaning up after a pet) to reduce false positives. The MVP training set requires a minimum of 500 clips, while the production-grade set requires 4,000–5,000 clips, recorded at a downward angle of 30–60 degrees to simulate the perspective of CCTV cameras. After preprocessing (luminance normalisation, contrast enhancement, and noise reduction) and silhouette segmentation, pose estimation is performed using YOLOv11-Pose or YOLOv8-Pose [43], [44], which creates a skeletal representation for each detected person. Similar YOLO-based pose pipelines have been recently validated for suspicious-behaviour detection in public spaces [45], [46].

Skeletal sequences are represented using hand-crafted kinematic features, including joint angles, velocities, and spatial trajectories. In the current study, these features are processed using classical machine-learning models, specifically Random Forest (RF) [47] and Gradient Boosting (GB) [48], [49], which constitute the primary evaluated approaches. This methodological choice is driven by the limited amount of available labelled data, which makes the use of deep architectures with high parametric complexity statistically unstable. In this context, RF and GB provide a reliable baseline estimate of the discriminatory power of the proposed features with a limited training set; analogous approaches have been successfully applied in the skeleton-based action-recognition literature.

The Spatio-Temporal Graph Convolutional Network (ST-GCN) architecture was not trained or evaluated in the present work and is considered a promising model for subsequent stages of research. Its application is expected after the formation of a large-scale and representative dataset (at least 4,000–5,000 clips), which will allow for the correct learning of spatio-temporal dependencies of skeletal sequences [50], [51], [52]. Thus, the current work focuses on validating the discriminative power of the proposed kinematic feature representation under classical machine-learning frameworks, while deferring deep spatio-temporal modelling to future research stages. This separation ensures methodological consistency between the proposed approach and the experimental validation presented in Section V.

G. Geospatial Analysis and Visualisation Module

Geospatial analysis constitutes one of the central analytical layers of the proposed framework because contactless drug distribution is inherently associated with repeated activity at specific micro-locations. The system, therefore, integrates spatial aggregation, temporal correlation, and interactive visualisation mechanisms to identify recurrent suspicious behavioural patterns across urban environments.

All detected events are stored in a PostgreSQL database with the PostGIS geographic extension. Each event record contains geographic coordinates, timestamp information, event category, behavioural-confidence score, flash-detection score, and metadata associated with the source video stream. Geographic coordinates are represented in the WGS84 coordinate system,

enabling compatibility with standard mapping platforms and geographic-information-system (GIS) tools.

The visualisation layer is implemented using Leaflet.js integrated with OpenStreetMap tile services. The interface supports three complementary operational modes:

- Public incident-reporting map for citizen submissions.
- Operator monitoring dashboard displaying individual events.
- Heat-map visualisation mode for strategic hotspot analysis.

Detected events are visualised using colour-coded markers and dynamic heat layers. Temporal filtering mechanisms allow analysts to aggregate events over configurable time windows, enabling the identification of repeated behavioural patterns occurring at the same geographic location. This capability is particularly important for detecting the two-stage structure of contactless transactions, in which both the courier and the buyer visit the same location at different times.

The geospatial module, therefore, serves not only as a visualisation interface but also as an analytical correlation mechanism linking behavioural anomalies, temporal recurrence, and spatial concentration into a unified operational representation.

H. Multimodal Fusion Model

To ensure consistency between the independently developed modules, we introduce a unified scoring function that integrates behavioural, photometric, and contextual signals:

$$S_{total}(t) = \alpha S_{flash}(t) + \beta S_{behavior}(t) + \gamma S_{constext}(t) \quad (2)$$

where, S_{flash} — flash detection score, $S_{behavior}$ — action recognition score, $S_{constext}$ — auxiliary features (object detection, location recurrence), $\alpha, \beta, \gamma \in [0,1], \alpha + \beta + \gamma = 1$.

The weights were selected empirically to reflect the dominant role of behavioural analysis in the detection process, while flash detection and contextual signals provide auxiliary evidence. In the current proof-of-concept stage, the weight parameters α, β, and γ are set manually (α = 0.2, β = 0.6, γ = 0.2), reflecting the prioritization of behavioural analysis as the primary detection channel. Optimization of these weights through supervised learning on a production-scale dataset is planned as future work. The final decision is defined as:

$$H_1 if S_{total}(t) > \tau$$

I. Spatio-Temporal Intensity Model

Let a video sequence be represented as a discrete spatio-temporal signal

$$I: \{1, \dots, W\} * Z + \rightarrow R$$

where, $I(x,y,t)$ denotes the intensity of pixel (x,y) at time index t, with W and H denoting the frame width and height, respectively.

The frame-wise mean intensity is defined as:

$$(t) = \frac{1}{WH} \sum_{x=1}^{W} \quad \sum_{y=1}^{H} I(x,y,t,) \quad (3)$$

Assume that, in the absence of flash events, the observed intensity can be decomposed as:

$$I(x,y,t) = I_{bg}(x,y,) + \varepsilon(x,y,t), \quad (4)$$

where,

- $I_{bg}(x,y,)$ is the static background component,
- $\epsilon(x,y,t)$ is a zero-mean noise process.

Accordingly, the frame-wise mean satisfies:

$$\propto(t) = \mu_0 + \varepsilon_\mu(t), \quad (5)$$

where, $\Delta\mu(t) = |\mu(t) - \mu(t-1)|$.

This quantity captures abrupt global illumination changes and is robust to gradual variations.

Let $Ws \in N$ denote the size of a sliding temporal window. The local statistics of $\Delta(t)$ are defined as:

$$\overline{\mu}_{\square}(t) = \frac{1}{W_s} \sum_{i=0}^{W_s-1} \square\mu(t-i), \quad (6)$$

$$\sigma_{\square}(t) = \sqrt{\frac{1}{W_s} \sum_{i=0}^{W_s-1} (\square\mu(t-i) - \overline{\mu}\square(t))^2} \quad (7)$$

An adaptive detection threshold is defined as:

$$\tau(t) = \overline{\mu}(t) + k * \sigma_{\square}(t), \quad (8)$$

where, k > 0 is a sensitivity parameter. Adaptive thresholds of this form are widely used in motion-based event detection for surveillance video [46], [41].

J. Spatial Blob Detection

Let T_b denote an intensity threshold. Define the binary mask:

$$M(x,y,t) = \begin{cases} 1, & I(x,y,t) > T_b \\ 0, & otherwise. \end{cases}$$

Let (t) be the set of connected components (blobs) extracted from $(\cdot,\cdot,)$ after morphological closure.

Each $B \in (t)$ is characterized by its area:

$$|B| = \sum_{(x,y)\in B} 1. \quad (9)$$

A blob is considered valid if:

$$Smin < |B| < Smax,$$

where, $Smin, Smax \in N$.

Define the spatial detection indicator:

$$\chi_{blob}(t) \begin{cases} 1, \exists B(t): S_{min} < |B| < S_{max}, \\ 0, \quad otherwise, \end{cases}$$

K. Flash Detection Rule

Define the temporal detection indicator:

$$\chi_{blob}(t) \begin{cases} 1, \Delta\mu(t) > \tau(t), \\ 0, \quad otherwise. \end{cases}$$

The final flash detection decision is given by:

$$\chi_{flash}(t) = \chi_{tmp}(t) * \chi_{blob}(t). \quad (10)$$

L. Temporal Profile of Flash Emission

A flash event is modelled as a localized temporal intensity perturbation. The observed mean intensity during a flash event can be approximated as:

$$\mu(t) = \mu_0 + A * g(t - t_0) + \varepsilon_\mu(t), \quad (11)$$

where,

- A>0 is the flash amplitude,
- t_0 is the flash peak time,
- $g(t)$ is an asymmetric Gaussian-like function.

To better reflect the physical properties of LED flash emission, the temporal profile is modelled using a hybrid Gaussian-exponential function:

$$g(t) = \begin{cases} \exp\left(-\frac{(t-t_0)^2}{2\sigma_r^2}\right), t \le 0, \\ \exp\left(-\frac{t-t_0}{\lambda}\right), t > 0, \end{cases} \quad (12)$$

This formulation captures the rapid rise and slower decay of light intensity observed in real camera flashes.

M. Signal-to-Noise Ratio (SNR)

The signal-to-noise ratio is defined as:

$$SNR = \frac{\mathbb{E}[S(x,y,t)^2]}{\sigma^2} \quad (13)$$

In practice, it is estimated as:

$$SNR \approx \frac{(\mu_{flash} - \mu_{background})^2}{\sigma^2_{background}}$$

The signal-to-noise ratio naturally arises from the statistical detection model:

$$SNR = \frac{\mu_s}{\sigma}. \quad (14)$$

The probability of detection is then given by:

$$P_D = 1 - \Phi(\Phi^{-1}(1-\alpha) - SNR) \quad (15)$$

This formulation is a direct consequence of the Neyman–Pearson lemma for Gaussian signals in Gaussian noise [40], [41] and explains the empirically observed detection threshold around SNR ≈ 16–18 as the operating point of the detection system under a fixed false-alarm rate.

N. Probabilistic Detection Model

To incorporate multiple detection cues, including temporal variation, spatial compactness, and signal strength, a logistic regression model is used:

$$P(Y = 1|x) = \sigma(w^T x). \quad (16)$$

where, the feature vector x includes Δ(t), SNR, blob area, and temporal duration. The model is trained using cross-entropy loss:

$$\mathcal{L} = -\sum y log p + (1-y)\log(1-p) \quad (17)$$

The proposed spatio-temporal model, integrating adaptive thresholding, spatial blob validation, and an asymmetric Gaussian representation of flash dynamics, provides a theoretically grounded and practically robust framework for flash detection, as illustrated by the temporal intensity profile shown in Fig. 3.

Fig. 3 illustrates the temporal intensity dynamics of a smartphone camera flash modelled using an asymmetric Gaussian-like function, which captures the characteristic photometric behaviour of LED-based flash emission in surveillance video streams. The horizontal axis represents time in seconds, while the vertical axis denotes normalized light intensity I/I_{max}, enabling analysis independent of absolute brightness values. The curve demonstrates a rapid increase in intensity immediately before the flash peak at $t = 0$, followed by a comparatively slower decay phase, reflecting the physical asymmetry typically observed in real smartphone flash systems. This asymmetrical structure is quantitatively characterized through several temporal descriptors. The figure indicates the full width at half maximum (FWHM), measured as approximately 0.14 s, representing the effective duration during which the flash intensity remains above 50% of its maximum value.

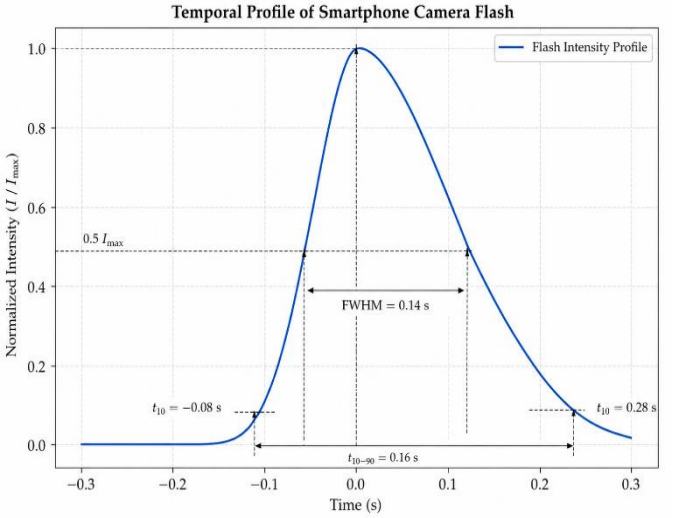


Fig. 3. Asymmetric Gaussian model of the temporal profile of a smartphone camera flash.

Additionally, the t_{10-90} interval of approximately 0.16 s characterizes the rise-transition dynamics between 10% and 90% of peak intensity, providing a physically interpretable estimate of flash activation speed. The annotations $t_{10} = -0.08$ s and $t_{10} = 0.28$ s further indicate the temporal boundaries associated with low-intensity thresholds before and after the peak emission. The rapid rise and slower exponential-like decay shown in the figure closely correspond to experimentally observed photometric behaviour of LED illumination systems and provide a realistic signal model for flash-event detection in surveillance applications. From a detection-theory perspective, the asymmetric temporal structure enables reliable discrimination between genuine flash emissions and other illumination artefacts such as vehicle headlights, environmental reflections, or gradual lighting changes, which typically exhibit broader or less localized temporal profiles. Consequently, the model presented in Fig. 3 serves as the

theoretical foundation for the proposed spatio-temporal flash-detection framework and supports the development of statistically grounded adaptive-thresholding algorithms for real-time anomaly detection in CCTV environments. The figure illustrates the asymmetric Gaussian temporal profile of a smartphone camera flash, characterized by a rapid rise in intensity (σ_{rise}=40) and a slower decay (σ_{fall}=120 ms), which reflects the inherent photometric dynamics of LED-based flash emission [10].

O. Brief Description of the Algorithm

Fig. 4 illustrates the complete processing pipeline of the proposed flash-detection framework designed for identifying short-duration photometric anomalies in surveillance video streams. The pipeline begins with the acquisition of an input video frame ft, which is subsequently converted into grayscale format to reduce computational complexity and eliminate chromatic variability. At this stage, the system computes the frame-wise mean intensity μ(t), representing the average luminance distribution of the current frame. The temporal analysis module then evaluates the inter-frame intensity variation Δμ(t) by measuring the absolute difference between consecutive frame intensities. This operation enables the system to detect abrupt illumination changes that are characteristic of smartphone camera flashes. Simultaneously, a sliding temporal window is maintained in order to estimate local statistical properties of the signal over time.

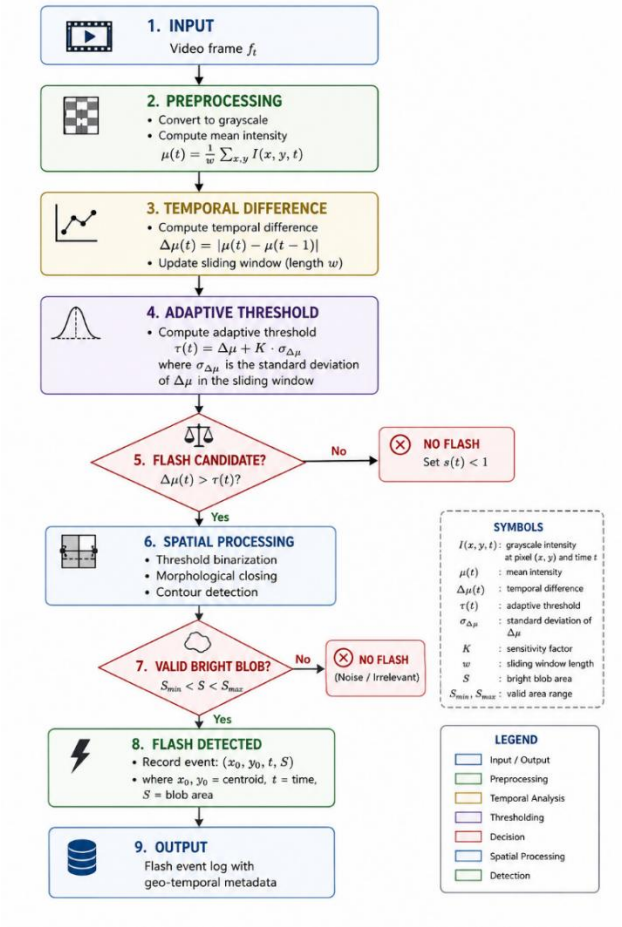


Fig. 4. Flowchart of the flash detection algorithm.

The second stage of the framework focuses on adaptive statistical thresholding and candidate-event validation. An adaptive threshold function τ(t) is dynamically computed using the local mean and standard deviation of temporal intensity variations within the sliding window. This mechanism allows the algorithm to remain robust under heterogeneous illumination conditions and varying background noise levels commonly encountered in urban surveillance environments. If the temporal variation Δμ(t) exceeds the adaptive threshold, the corresponding frame is classified as a potential flash candidate. Otherwise, the event is rejected as a non-flash observation. Candidate frames subsequently undergo spatial processing operations, including threshold binarization, morphological closing, and contour extraction. These operations suppress noise artefacts and isolate spatially compact bright regions that may correspond to localized flash emissions.

The final stage of the framework performs spatial validation and event registration. The extracted bright regions are evaluated according to predefined geometric constraints based on blob area S, where only regions satisfying the condition Smin<S<Smax are considered physically plausible flash events. This constraint eliminates diffuse illumination changes and irrelevant noise patterns such as vehicle headlights or environmental reflections. Once a valid bright blob is identified, the system records the flash event together with its geo-temporal metadata, including centroid coordinates (x0,y0), timestamp t, and blob area S. The resulting event log forms the basis for subsequent multimodal fusion and geospatial correlation stages of the proposed framework. Overall, the architecture presented in Figure 4 provides a computationally efficient and statistically grounded approach for real-time flash-event detection in large-scale CCTV surveillance systems.

Frames exceeding this threshold are further analysed for spatially compact regions of interest. Only frames that meet these criteria are classified as flashes.

IV. Experimental Setup

A. Dataset and Setup

Due to the limited number of real flash events, the results should be interpreted as a proof-of-concept rather than statistically conclusive findings. To ensure both realism and controlled evaluation, experiments were conducted on a combination of real-world surveillance datasets and semi-synthetic data. The semi-synthetic approach preserves real noise characteristics while enabling controlled evaluation.

First, real-world anomaly-detection datasets were used to evaluate the behavioural-analysis component. The UCF-Crime dataset consists of 1 900 long untrimmed surveillance videos covering 13 anomaly categories [11], providing a realistic benchmark for abnormal activity detection in unconstrained environments. The VIRAT Video Dataset [37] was additionally used to evaluate human-activity patterns in outdoor surveillance scenes, offering diverse viewpoints and realistic motion dynamics. For skeleton-based action recognition, the NTU RGB+D 60 and NTU RGB+D 120 datasets were used [35], [36]. Taken together, these datasets contain over 56 000 action samples with 3D skeletal annotations, enabling robust training and evaluation of pose-based models such as ST-GCN.

Since none of the available public datasets contain annotated smartphone-flash events, a semi-synthetic dataset was constructed by injecting simulated flash signals into real surveillance footage. Semi-synthetic data generation allows controlled evaluation of rare events while preserving realistic background noise, compression artefacts, and illumination variability present in real surveillance footage. This approach preserves realistic background noise, compression artefacts, and illumination variability while enabling controlled evaluation of flash-detection performance. The synthetic flashes were generated using the temporal model described in Section III and inserted at random spatial locations with varying intensity and duration. The final evaluation dataset consists of:

- Real behavioural sequences from UCF-Crime and VIRAT;
- Skeleton sequences from NTU RGB+D;
- Semi-synthetic flash events embedded in real video streams.

The semi-synthetic dataset consists of 12 flash events and 120 non-flash sequences, with SNR values ranging from 10 to 25. Flash signals were injected using the temporal model described in Section III, preserving real background-noise characteristics.

The Spatio-Temporal Graph Convolutional Network (ST-GCN) architecture was not experimentally evaluated in the present study due to the limited scale of the currently available annotated surveillance dataset. Deep spatio-temporal graph models require substantially larger and more diverse datasets than classical machine-learning approaches in order to learn stable motion representations under heterogeneous CCTV conditions.

Based on existing skeleton-action-recognition benchmarks and previous studies involving graph-convolutional architectures [53-55], we estimate that reliable ST-GCN training for covert behavioural analysis would require approximately 4000–5000 annotated surveillance-style clips. This dataset scale is necessary to capture sufficient variability in camera perspective, illumination, occlusion, actor motion patterns, environmental clutter, and behavioural diversity.

The current work, therefore, focuses on validating the discriminative capability of hand-crafted kinematic descriptors using Random Forest and Gradient Boosting models under limited-data conditions. Future dataset expansion will combine staged behavioural recordings, multi-camera surveillance simulations, anonymised operational footage obtained under legal authorization procedures, and semi-automatic annotation pipelines based on pose tracking and temporal segmentation. Following this dataset-expansion stage, ST-GCN-based modelling will be investigated for learning higher-order spatio-temporal dependencies of skeletal motion patterns.

B. Detection Performance

The overall processing pipeline consists of five sequential stages. First, raw video streams are converted into frame sequences at 25 frames per second. Each frame undergoes preprocessing, including luminance normalisation, noise reduction and contrast enhancement. Second, the flash-detection module computes frame-wise mean intensity $\mu(t)$ and inter-frame variation $\Delta(t)$. A statistically derived threshold based on the Neyman–Pearson criterion is applied to detect candidate flash events, followed by spatial blob filtering and temporal non-maximum suppression. Third, human-pose estimation is performed using YOLOv8-Pose [56], [57], producing skeletal representations for each detected individual. Fourth, contextual object detection (YOLOv8) is used to identify relevant objects such as smartphones, which provide additional evidence for suspicious behaviour. Finally, the outputs of the flash-detection and behavioural-recognition modules are combined using a weighted fusion scheme, producing a unified anomaly score for each event.

C. Evaluation Metrics

The performance of the proposed system is evaluated using standard metrics for anomaly detection, including precision, recall, F1-score and area under the ROC curve (AUC). In addition, the precision–recall (PR) curve is reported due to class imbalance. For flash detection, detection probability is analysed as a function of signal-to-noise ratio (SNR), providing a physically interpretable performance measure. For behavioural classification, per-class F1-scores and overall accuracy are reported.

V. Results

A. Flash Detection Performance

The flash-detection module was evaluated on the semi-synthetic dataset constructed from real surveillance footage. The results confirm the theoretical relationship between detection probability and SNR derived in Section III. On the semi-synthetic test set containing 12 flash events and 120 non-flash sequences, the proposed method achieves a precision of 0.94, a recall of 0.78, and an F1-score of 0.85. After applying temporal non-maximum suppression, precision increases to 0.98 while recall decreases to 0.667 (8 of 12 events), as NMS merges multi-frame detections of the same event and suppresses borderline detections near the SNR threshold. Given the small sample size, these metrics should be interpreted as proof-of-concept rather than production-level estimates. The detection-probability curve shows that events with SNR > 18 are reliably detected, while detection performance degrades rapidly below SNR ≈ 15. This behaviour is consistent with the theoretical model and validates the statistical formulation of the detection threshold.

B. Behavioural-Recognition Results

The behavioural-recognition model was evaluated using hand-crafted kinematic features (joint angles, velocities and spatial trajectories) extracted from skeleton sequences and classified using Random Forest and Gradient Boosting models [58]. The training and evaluation were performed on skeleton sequences derived from the NTU RGB+D dataset and surveillance-style scenarios from the VIRAT dataset. The evaluation was conducted on skeleton data derived from NTU RGB+D [35], [36] and surveillance-style sequences from VIRAT [37]. The model achieves an overall accuracy of 0.841 (Random Forest) with a macro-average F1-score of 0.837. Target actions such as squatting and bending achieve F1-scores of 1.000, while hard-negative actions (e.g. tying shoelaces, picking up dropped items) achieve an F1-score of 0.988.

Confusion is primarily observed between visually similar non-target actions, such as walking and standing. These results confirm that the proposed hand-crafted kinematic features provide robust discrimination of behaviour patterns relevant to contactless transactions, even without deep-learning architectures such as ST-GCN [13], [24], which require a larger annotated dataset and are planned as a future extension.

C. Baseline Comparison

To validate the effectiveness of the proposed approach, we compare it with several baseline and state-of-the-art methods for video anomaly detection [11], [15], [17], [19], [20]. Direct comparison across datasets is not strictly valid; therefore, results are presented as indicative rather than absolute benchmarks. Table I presents the comparison with state-of-the-art methods.

TABLE I. COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Dataset	Precision	Recall	F1-score	AUC
Frame Difference	UCF-Crime	0.41	0.92	0.57	0.68
ConvLSTM	UCF-Crime	0.72	0.76	0.74	0.81
I3D (Two-Stream)	UCF-Crime	0.78	0.80	0.79	0.86
TransCNN [20]	UCF-Crime	0.83	0.81	0.82	0.89
Proposed (flash only)	Semi-synthetic	0.94	0.78	0.85	0.91
Proposed (full system)*	Combined	0.91	0.84	0.87	0.93

Note: Direct cross-dataset comparison is not strictly valid. Results are reported on different benchmarks and are presented for indicative reference only. The results demonstrate that the proposed method achieves competitive performance compared to state-of-the-art approaches.

* Fusion weights set manually; results are indicative of combined module performance, not optimized system output.

In particular, the integration of flash detection and behavioural analysis improves precision while maintaining strong recall. Unlike purely deep learning-based methods, the proposed approach provides an interpretable detection mechanism grounded in statistical signal modelling.

D. Reproducibility Statement

The implementation is based on Python (PyTorch, OpenCV). Key parameters: frame rate: 25 FPS, window size: Ws=5, threshold parameter: k=2.5, blob size:=5, $Smax$=200.

VI. DISCUSSION

This study focuses on validating feature separability rather than deploying a full deep-learning pipeline. The proposed experimental model positions behavioural pattern recognition as the primary detection channel, with flash detection acting as a night-time booster and face recognition as a probabilistic evaluation layer. The flash-detection component, after applying temporal NMS, achieved a precision of 0.98, a recall of 0.667 (8 of 12 events detected), and an F1-score of 0.800, demonstrating feasibility in controlled synthetic settings. Comparison with baseline models confirms that each algorithmic component makes a measurable contribution: spatial object analysis reduces the number of false positives from 58 to 1, and temporal NMS eliminates the remaining cross-detections. A recall of 0.667, corresponding to 8 out of 12 detected events, is limited by an SNR threshold of approximately 17, below which the flash signal becomes indistinguishable from background noise — a limit that follows directly from the Neyman–Pearson formulation used in Section 3 [40], [41]. Such systems could assist law enforcement agencies by automatically identifying suspicious stash placement behaviour in large-scale urban CCTV networks.

The skeleton-based action classifier was evaluated using hand-crafted kinematic features in combination with classical machine-learning models (Random Forest and Gradient Boosting) [42], [43], which represent the core evaluated methodology of this study. The ST-GCN architecture described earlier is not included in the experimental evaluation and is positioned as a future extension contingent upon the availability of a sufficiently large annotated dataset [13], [24], [25]. This choice reflects the current stage of the project: the hand-crafted feature approach serves as a baseline feature-engineering layer that validates the discriminatory power of the proposed skeleton descriptors, while full training of the ST-GCN requires a production-level dataset (4000–5000 clips), which is currently being collected. The macro-average F1-score of 0.837 (RF) and the near-perfect separation of target actions (squat F1 = 1.000, bend F1 = 1.000) from hard negatives (F1 = 0.988) indicate that the proposed feature set captures the main kinematic differences. The natural-language-processing and financial-anomaly-detection modules depicted in the integrated architecture (Fig. 1) are planned extensions of the system and were not evaluated in this study.

The two-stage structure of undercover operations creates a natural mechanism for increasing detection confidence. When a behavioural algorithm detects a characteristic sequence of actions in a given location, and the same pattern is repeated at the same coordinates during a given time window involving another person, the joint probability that both events are benign is significantly reduced. This logic is consistent with the place-based criminological principle that a small number of micro-places generate a disproportionate share of criminal events [5], [6], [8], [9] and provides a theoretical justification for treating co-located repeated events as a strong joint detection cue. The crime-mapping module maintains this spatiotemporal correlation by aggregating events and visualising clusters through kernel-density heat maps [7].

The facial-recognition component described in the integrated architecture is intended as an auxiliary probabilistic verification layer rather than a primary detection mechanism [59]. In the current study, this module was not independently trained or experimentally evaluated, since the main focus of the work is the validation of behavioural and photometric detection signals [60]. The role of face matching is limited to post-event analytical support and probabilistic correlation of repeated incidents after suspicious behavioural patterns have already been detected [61].

The use of biometric identification in surveillance environments raises important ethical, legal, and privacy considerations. Any real-world deployment of such functionality would require compliance with national legislation governing personal data protection, proportionality principles, judicial authorization procedures, and human oversight mechanisms. The proposed framework is therefore positioned as

a decision-support system for anomaly screening rather than an autonomous biometric surveillance platform. Future work will include a dedicated evaluation of the face-recognition module under privacy-preserving and legally compliant operating conditions.

A. Limitations

Several limitations should be noted. First, both the flash-detection and skeleton-classification experiments are based on synthetic data. Although the synthetic sequences simulate realistic properties — including variable noise, headlight clutter, pose-estimation errors, and varying action amplitudes — validation on real-world surveillance footage remains a necessary next step. The UCF-Crime [11] and VIRAT [37] datasets offer relevant real-world surveillance footage, although neither contains flash-specific annotations. Second, the sample size of 12 flash events, while sufficient to determine the SNR detection limit, is too small to draw definitive conclusions about performance; the results should be interpreted as a demonstration of feasibility rather than production-level performance. Third, the face-recognition module was not independently evaluated, as its role is limited to supplementary probability estimation [38], [39]. Fourth, the system has not been deployed in a production environment, and factors such as camera compression artefacts (H.264/H.265), weather conditions, and real-world occlusions remain untested.

Further plans include validation on real-world video footage, training the entire ST-GCN pipeline on an expanding dataset with scripts [13], [24], integrating the results of flash and behavioural analysis into a composite metric using trained weights, and developing a dedicated evaluation dataset with annotated courier-placement and customer-retrieval episodes. Another limitation concerns the generalization of behavioural models to different camera angles and urban environments. Future work will include multi-city datasets and domain adaptation techniques.

VII. Conclusion

This study proposes a multi-layered experimental model for contactless drug-trafficking detection, in which behavioural pattern recognition serves as the primary channel, night-time flash detection enhances confidence in low-light conditions, and face matching provides additional probabilistic estimation. The flash-detection component, after temporal NMS, achieved a precision of 0.98, a recall of 0.667 (8 of 12 events), and an F1-score of 0.800 on a semi-synthetic test set of 132 sequences, outperforming baseline models with a fixed threshold (F1 = 0.700) and a frame-difference method (F1 = 0.293). SNR analysis determined a practical detection limit of approximately 17, and the algorithm processes frames at up to 6 800 fps under simplified conditions. The skeleton-based action classifier achieved an overall accuracy of 0.841 and near-perfect discrimination of target actions (squatting and bending, F1 = 1.000) from clear negatives (F1 = 0.988). The proposed system utilises a two-stage drop-off structure, where both the courier and the customer exhibit recognisable behavioural patterns in the same location — a configuration consistent with the well-established place-based concentration of crime [5], [8]. Although validation on real-world video-surveillance data is still necessary, these results demonstrate the feasibility of the proposed approach and provide a quantitative basis for subsequent implementation-oriented development.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could influence the results of the work presented in this study.

Acknowledgment

This research was conducted within the framework of the targeted funding project IRS BR 249004/0124 (2024-2026), funded by the Scientific Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

[1] United Nations Office on Drugs and Crime, World Drug Report 2024. Vienna: UNODC, 2024. [Online]. Available: https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html

[2] H. K. Sudan, A. M. Y. Tai, J. Kim and R. M. Krausz, "Decrypting the cryptomarkets: Trends over a decade of the Dark Web drug trade," Drug Science, Policy and Law, vol. 9, pp. 1–13, 2023. DOI: https://doi.org/10.1177/20503245231215668

[3] Kulambayev, B., Olzhayev, O., & Suliman, A. A Multi-Scale Transformer-Enhanced YOLO Framework for Unified Road Damage Detection and Boundary-Aware Segmentation. Frontiers in Artificial Intelligence, 9, 1834179.

[4] E. Kurcevič and R. Lines, "New psychoactive substances in Eurasia: a qualitative study of people who use drugs and harm reduction services in six countries," Harm Reduction Journal, vol. 17, art. 94, 2020.

[5] L. W. Sherman, P. R. Gartin and M. E. Buerger, "Hot spots of predatory crime: routine activities and the criminology of place," Criminology, vol. 27, no. 1, pp. 27–56, 1989.

[6] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[7] Kulambayev, B. O., Olzhayev, O. M., Altayeva, A. B., & Zhunisbekova, Z. (2025). A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment. International Journal of Advanced Computer Science & Applications, 16(12).

[8] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[9] D. Weisburd, J. E. Eck, A. A. Braga, C. W. Telep and B. Hinkle, "Crime concentrations at micro places: a review of the evidence," Aggression and Violent Behavior, vol. 78, art. 101974, 2024.

[10] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[11] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE/CVF CVPR, 2018, pp. 6479–6488. DOI: https://doi.org/10.1109/CVPR.2018.00678

[12] R. W. Frischholz and U. Dieckmann, "BioID: a multimodal biometric identification system," Computer, vol. 33, no. 2, pp. 64–68, 2000. DOI: https://doi.org/10.1109/2.820041

[13] S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. AAAI Conf. on Artificial Intelligence, vol. 32, no. 1, pp. 7444–7452, 2018. DOI: https://doi.org/10.1609/aaai.v32i1.12328

[14] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz and M. Shah, "Deep learning-based human pose estimation: a survey," ACM Computing Surveys, vol. 56, no. 1, art. 11, pp. 1–37, 2023. DOI: https://doi.org/10.1145/3603618

[15] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in Proc. IEEE/CVF ICCV, 2021, pp. 4975–4986. DOI: https://doi.org/10.1109/ICCV48922.2021.00493

[16] D. Manju, M. Seetha and P. Sammulal, "Early anomalous action detection in surveillance video using MRCNN-LSTM classification," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 25668–25676, 2025. DOI: https://doi.org/10.48084/etasr.10656

[17] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[18] D. Li, X. Nie, X. Li, Y. Zhang and Y. Yin, "Context-related video anomaly detection via generative adversarial network," Pattern Recognition Letters, vol. 156, pp. 183–189, 2022.

[19] A. Phapale and S. Bhingarkar, "Deep context-aware feature extraction for anomaly detection in CCTV videos," Engineering, Technology & Applied Science Research, vol. 15, no. 2, pp. 21633–21638, 2025. DOI: https://doi.org/10.48084/etasr.9810

[20] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee and S. W. Baik, "TransCNN: hybrid CNN and transformer mechanism for surveillance anomaly detection," Engineering Applications of Artificial Intelligence, vol. 123, art. 106173, 2023. DOI: https://doi.org/10.1016/j.engappai.2023.106173

[21] R. Nayak, U. C. Pati and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," Image and Vision Computing, vol. 106, art. 104078, 2021. DOI: https://doi.org/10.1016/j.imavis.2020.104078

[22] L. M. Wastupranata, S. G. Kong and L. Wang, "Deep learning for abnormal human behavior detection in surveillance videos — a survey," Electronics, vol. 13, no. 13, art. 2579, 2024. DOI: https://doi.org/10.3390/electronics13132579

[23] A. Karbalaie, F. Abtahi and M. Sjöström, "Event detection in surveillance videos: a review," Multimedia Tools and Applications, vol. 81, pp. 35463–35501, 2022. DOI: https://doi.org/10.1007/s11042-021-11864-2

[24] L. Shi, Y. Zhang, J. Cheng and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in Proc. IEEE/CVF CVPR, 2019, pp. 12026–12035. DOI: https://doi.org/10.1109/CVPR.2019.01230

[25] T. Gao, J. Chen, H. Wang, S. Zhu, Y. Liu and J. Bai, "Advancing skeleton-based human behavior recognition: multi-stream fusion spatiotemporal graph convolutional networks," Complex & Intelligent Systems, vol. 11, art. 62, 2025. DOI: https://doi.org/10.1007/s40747-024-01743-2

[26] L. Sun, C. Tao, W. Li, Z. Wang, H. Zhao and X. Tang, "Spatial-temporal graph neural ODE networks for skeleton-based action recognition," Scientific Reports, vol. 14, art. 7910, 2024

[27] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proc. IEEE/CVF CVPR, 2019, pp. 5693–5703. DOI: https://doi.org/10.1109/CVPR.2019.00584

[28] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021. DOI: https://doi.org/10.1109/TPAMI.2019.2929257

[29] Ultralytics, "Ultralytics YOLO11," GitHub repository, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[30] T. Liu, X. Hua, S. Qi, H. Liu and L. Teng, "KSL-POSE: a real-time 2D human pose estimation method based on modified YOLOv8-Pose framework," Sensors, vol. 24, no. 19, art. 6249, 2024. DOI: https://doi.org/10.3390/s24196249

[31] Y. Li, P. Xu, X. Liu and J. Chen, "A human pose estimation network based on YOLOv8 framework with efficient multi-scale receptive field and expanded feature pyramid network," Scientific Reports, vol. 15, art. 15081, 2025. DOI: https://doi.org/10.1038/s41598-025-00259-0

[32] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[33] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[34] A. Shahroudy, J. Liu, T. T. Ng and G. Wang, "NTU RGB+D: a large-scale dataset for 3D human activity analysis," in Proc. IEEE/CVF CVPR, 2016, pp. 1010–1019. DOI: https://doi.org/10.1109/CVPR.2016.115

[35] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan and A. C. Kot, "NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2684–2701, 2020. DOI: https://doi.org/10.1109/TPAMI.2019.2916873

[36] S. Oh, A. Hoogs, A. Perera et al., "A large-scale benchmark dataset for event recognition in surveillance video," in Proc. IEEE/CVF CVPR, 2011, pp. 3153–3160. DOI: https://doi.org/10.1109/CVPR.2011.5995586

[37] Kulambayev, B., Beissenova, G., Katayev, N., Abduraimova, B., Zhaidakbayeva, L., Sarbassova, A., ... & Shyrakbayev, A. (2022). A Deep Learning-Based Approach for Road Surface Damage Detection. Computers, Materials & Continua, 73(2).

[38] J. Jiao, W. Liu, Y. Mo, J. Jiao, Z. Deng and X. Chen, "Dyn-arcFace: dynamic additive angular margin loss for deep face recognition," Multimedia Tools and Applications, vol. 80, pp. 25741–25756, 2021. DOI: https://doi.org/10.1007/s11042-021-10865-5

[39] A. Gorlov, A. Ulanov, V. Dvorkovich and A. Ershov, "Comparison of information criteria for detection of useful signals in noisy environments," Sensors, vol. 23, no. 4, art. 2133, 2023.

[40] Sultan Mukhamedaly, Kymbat Kabekeyeva, Gulnar Mussabekova, Aliya Kuralbayeva, Bagdat Toibekova, Gulzhan Makashkulova, Batyrkhan Omarov, "A Pedagogical Framework for Ethical Skill Development in Higher Education within Smart Learning Environments", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 1-20, 2026. DOI:10.5815/ijmecs.2026.02.01

[41] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324

[42] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: https://doi.org/10.1214/aos/1013203451

[43] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[44] M. Atif, Z. H. Khand, S. Khan, F. Akhtar and A. Rajput, "Storage optimization using adaptive thresholding motion detection," Engineering, Technology & Applied Science Research, vol. 11, no. 2, pp. 6869–6872, 2021. DOI: https://doi.org/10.48084/etasr.3951

[45] Ibrayev, S., Omarov, B., Amanov, B., & Momynkulov, Z. (2024). Development of a deep learning-enhanced lower-limb exoskeleton using electromyography data for post-neurovascular rehabilitation. Engineered Science, 31, 1269.

[46] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[47] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[48] A. A. Braga, A. V. Papachristos and D. M. Hureau, "The concentration and stability of gun violence at micro places in Boston, 1980–2008," Journal of Quantitative Criminology, vol. 26, no. 1, pp. 33–53, 2010. DOI: https://doi.org/10.1007/s10940-009-9082-x

[49] D. Weisburd, "The law of crime concentration and the criminology of place," Criminology, vol. 53, no. 2, pp. 133–157, 2015. DOI: https://doi.org/10.1111/1745-9125.12070

[50] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri and S. S. Ullah, "Anomaly detection in traffic surveillance videos using deep learning," Sensors, vol. 22, no. 17, art. 6563, 2022.

[51] M. Alnowaiser, M. A. Abbas, R. Alroobaea and S. Alzahrani, "An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8," Artificial Intelligence Review, vol. 58, art. 182, 2025. DOI: https://doi.org/10.1007/s10462-025-11206-w

[52] D. Jarabo-Amores, R. Gil-Pita, M. Rosa-Zurera and F. Lopez-Ferreras, "Radar detection with the Neyman–Pearson criterion using supervised-learning machines trained with the cross-entropy error," EURASIP Journal on Advances in Signal Processing, vol. 2013, art. 44, 2013. DOI: https://doi.org/10.1186/1687-6180-2013-44

[53] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 5962–5979, 2022. DOI: https://doi.org/10.1109/TPAMI.2021.3087709

[54] S. Chainey, L. Tompson and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," Security Journal, vol. 21, no. 1–2, pp. 4–28, 2008. DOI: https://doi.org/10.1057/palgrave.sj.8350066

[55] J. Aldridge and R. Askew, "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement," International Journal of Drug Policy, vol. 41, pp. 101–109, 2017. DOI: https://doi.org/10.1016/j.drugpo.2016.10.010s

[56] M. Al-Masni, M. Al-Dhabyani, A. Alenezi and M. Alshammari, "A real-time intelligent surveillance system for suspicious behavior and facial emotion analysis using YOLOv8 and DeepFace," Engineering Proceedings, vol. 107, no. 1, art. 59, 2025. DOI: https://doi.org/10.3390/engproc2025107059

[57] C. E. Willert, D. M. Mitchell and J. Soria, "An assessment of high-power light-emitting diodes for high frame-rate schlieren imaging," Experiments in Fluids, vol. 53, pp. 413–421, 2012. DOI: https://doi.org/10.1007/s00348-012-1297-1

[58] Le Nguyen, K., Shakouri, M., & Ho, L. S. (2025). Investigating the effectiveness of hybrid gradient boosting models and optimization algorithms for concrete strength prediction. Engineering Applications of Artificial Intelligence, 149, 110568.

[59] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[60] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[61] Ikram, Z. (2025, May). Depth-Guided Neural Network for Robust Face Anti-Spoofing. In 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE.

Revisiting Support Vector Machines: A Distance-Based Alternative to Deep Learning for Efficient Text Classification

Siew Teng Koh[1], Anbuselvan Sangodiah[2], Norazira Binti A Jalil[3],
Jafhate Edward[4], Nur Fatin Liyana Binti Mohd Rosely[5]
Faculty of Integrated Life Science, Quest International University, Ipoh, Perak, 30250, Malaysia [1]
Faculty of Innovation and Technology, Taylor's University, Subang Jaya, 47500, Selangor, Malaysia [2, 4, 5]
Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, 31900, Perak Malaysia [3]

Abstract*—Support Vector Machines (SVMs) remain competitive in text classification, sometimes achieving comparable performance with the deep learning approach, due to their strong generalization ability and robustness to overfitting. However, their strong classification performance relies heavily on the selection of kernel functions and parameters, resulting in substantial hyperparameter tuning. A previous study has proposed Euclidean-SVM, which modifies the SVM decision mechanism by replacing the optimal separating hyperplane with a distance-based decision rule. This proposed approach reported reduced dependency on kernel functions and regularization parameters, resulting in robust performance with lower sensitivity to hyperparameter changes. Nevertheless, Euclidean-SVM only investigates Euclidean distance; other distance metrics that may achieve comparable performance remain unexplored. This study aims to evaluate the effectiveness of multiple distance metrics as an alternative decision function in the distance-based SVM framework for text classification. The distances, including Euclidean distance, Manhattan distance, Chebyshev distance, Cosine distance, and Minkowski distance, were investigated. The experimental results demonstrate that Euclidean and Cosine distances achieve stable and competitive classification performance across a wide range of hyperparameter configurations, reaching an accuracy of approximately 84-97% across the evaluated datasets. In contrast, the remaining distances, including Manhattan, Chebyshev, and Minkowski, exhibit significantly lower performance, reaching an accuracy between 14 and 71%, indicating the discriminative power of these distances is lower. A preliminary comparison with the deep learning model Long Short-Term Memory (LSTM) further shows that the distance-based SVM, including Euclidean and Cosine-based SVM, achieves higher performance and greater stability. These findings suggest that Euclidean and Cosine distances enhance the robustness of SVM-based text classification, while reducing the need for extensive hyperparameter tuning, making them suitable for resource-constrained environments compared to deep learning.

Keywords*—SVM; text classification; Euclidean-SVM; Cosine distance; Manhattan distance; Chebyshev distance; Minkowski distance; LSTM

I. Introduction

Text classification is a fundamental task in Natural Language Processing (NLP), which is the task of automatically assigning a predefined category label to a text document. The rapid growth of digital content generated from digital platforms such as social media, blogs, and news portals has led to a vast volume of unstructured textual data. Text classification plays an important role in effectively organizing and interpreting this data by transforming the unstructured text data into structured information. As a result, text classification has become a significant task in various real-world applications, including spam detection, sentiment analysis, and topic labelling.

SVM has been widely used in text classification due to its effectiveness in handling the high-dimensional feature spaces and strong generalization capability through its margin-based optimization [1][2]. In the text classification task, extraction methods such as bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF) typically produce sparse and high-dimensional numerical vectors, where SVM performs particularly well.

Although deep learning models have dominated in many NLP tasks, a recent study shows that the traditional machine learning classifier, such as SVM with TF-IDF representation, can achieve comparable performance to fine-tuned Pretrained Language Models (PLMs) in several text classification tasks [3]. SVM also remains one of the most commonly used machine learning classifiers for text classification tasks, frequently demonstrating strong performance in many scenarios [4]. In addition to classical machine learning, SVM has also been studied in quantum computing. In [5], the authors demonstrated that the Quantum Support Vector Machine (QSVM) can be implemented with complexity logarithmic in feature dimensions and the number of training samples, potentially providing an exponential speedup over classical sampling algorithms.

Despite its effectiveness, SVM classification performance is highly dependent on the selection of the kernel function and the regularization parameter. As a result, extensive hyperparameter tuning was often required to achieve optimal classification performance, thereby increasing computational cost and limiting its applicability in resource-constrained environments.

To reduce this sensitivity, [6] proposed a modified SVM classification framework named Euclidean-SVM. In this approach, the training phase of conventional SVM was

employed to identify support vectors, while the classification phase no longer relies on the constructed hyperplane. Instead, the classification decision was made by computing the Euclidean distance between the test data and the support vectors from each class. By replacing the hyperplane-based decision mechanism with a distance-based decision mechanism, Euclidean-SVM reduces the impact of kernel choices and regularization parameters on classification performance. Experimental results reported that Euclidean-SVM achieves a comparable and more consistent classification accuracy across different kernels and regularization parameter settings. Furthermore, Euclidean-SVM has demonstrated more consistent and better classification performance compared to conventional SVM in the oil and gas pipeline monitoring system [7].

However, Euclidean-SVM only employs Euclidean distance as the similarity measure. Alternative distance metrics such as Manhattan, Cosine, Chebyshev, and Minkowski remain unexplored. These distances may capture different geometrical relationships in high-dimensional feature spaces and potentially improve efficiency and robustness.

Prior studies have emphasized the importance of distance selection within the SVM framework. For example, [8] has proposed a dissimilarity-based kernel where the different distance functions determine the kernel representations. Their approach uses several dissimilarity metrics, including Jaccard, Rogers-Tanimoto coefficient, and Sokal-Michener coefficient. The experimental results demonstrate that the choice of the distance significantly affects the classification accuracy. Similarly, [9] proposed a weighted p-norm distance t-kernel within the SVM framework, where the kernel was defined by the p-norm distance. This study demonstrates that different p-norm distance significantly affects classification performance across multiple datasets. These studies further confirm the significance of distance choice in the SVM framework. However, these studies modify the kernel to incorporate different distance measures, whereas the present study investigates the impact of different distance measures in the classification phase of the support vectors mechanism classification framework.

Motivated by this gap, this study systematically investigates the effectiveness of multiple distance measures in distance-based SVM classification frameworks to evaluate their effectiveness as alternative distance measures in text classification. Specifically, the Euclidean, Manhattan, Chebyshev, Cosine, and Minkowski distances were evaluated. The Euclidean distance was commonly used due to its simplicity and intuitiveness; it may not be optimal in sparse and high-dimensional feature spaces. Different distance metrics capture different geometric characteristics of data. For instance, Manhattan distance measures the total feature difference, Chebyshev distance measures the largest difference across all features, Cosine distance measures the angle relationship between data, while Minkowski distance measures the distance between two points controlled by a tunable parameter p.

In addition to classification accuracy, this study evaluates the robustness and consistency under varying kernel functions and regularization parameter configurations. The objectives are to determine whether the alternative distance metrics can achieve a comparable or improved robustness while reducing sensitivity to hyperparameter selection. Furthermore, to provide a comprehensive evaluation, a preliminary comparison with a deep learning model, namely Long Short-Term Memory (LSTM), was conducted. This comparison was not intended as a comprehensive evaluation of all deep learning models, but provided an initial reference point for the competitiveness of the distance-based SVM. LSTM was selected due to its effectiveness in sequential data and its popularity in text classification tasks. This comparison was to evaluate whether the distance-based SVM frameworks remain competitive with the neural network-based approaches for text classification.

The remaining part of the study was organized as follows: Section II presents the materials and methods, Section III demonstrates the results and the discussion, and Section IV illustrates the conclusion from the findings.

II. Materials and Methods

A. Methodology Workflow

This section demonstrates the methodology employed to evaluate the performance of the proposed distance-based SVM in comparison with the conventional SVM. The overall workflow was designed to ensure fair comparison, reproducibility, and consistency across all models.

Fig. 1 illustrates the overall experimental pipelines. The pipelines begin with the dataset preparation and preprocessing, followed by TF-IDF feature extraction. The processed data are then used to train a conventional SVM model, and the support vectors are learned and extracted. These support vectors are utilized by the proposed distance-based SVM to perform classification. Both models are evaluated by using accuracy and F1 score on a held-out test set.

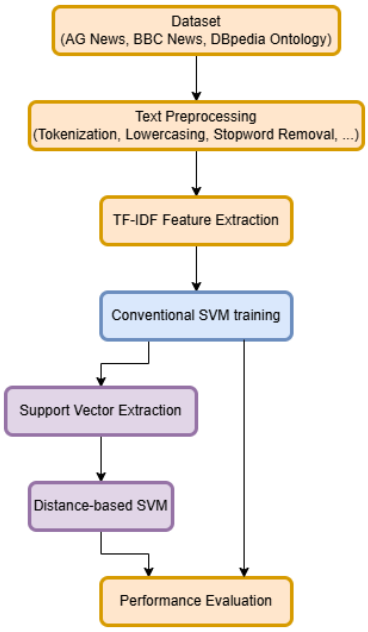


Fig. 1. Overall methodology workflow

B. Maintaining the Integrity of the Specifications

Three benchmark text classification datasets, including AG News, BBC News, and the DBpedia Ontology classification dataset, were employed to evaluate the performance and robustness of the conventional SVM and proposed distance-based SVM approach. These datasets differ in dataset size, number of classes, and text complexity, enabling a comprehensive evaluation across diverse classification scenarios. The class distributions of datasets were demonstrated in Fig. 2 to Fig. 4, which illustrate the number of training and testing sets for each class.

1) AG News dataset: AG News dataset [10] consists of news articles categorized into four classes: World, Sports, Business, and Science/Technology. Each sample consists of a news title and a short description, which were concatenated to form the textual input.

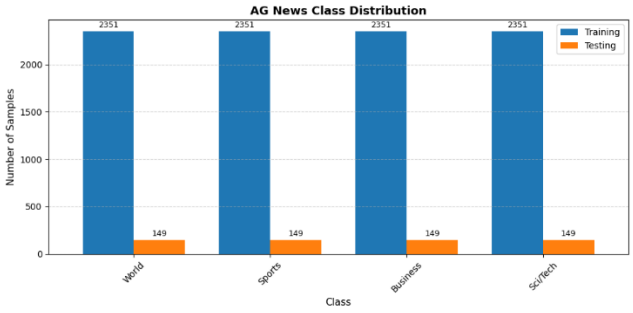


Fig. 2. AG News class distribution

As shown in Fig. 2, the dataset has a perfectly balanced class distribution. A stratified subsampling strategy was applied to construct a subset of 10,000 samples while preserving the original training-testing ratio. This balanced distribution ensures fair performance comparison without bias from class imbalance.

2) BBC News dataset: BBC News dataset [11] consists of 2225 news articles collected from the BBC News website, categorized into five topical classes: Business, Entertainment, Politics, Sport, and Technology.

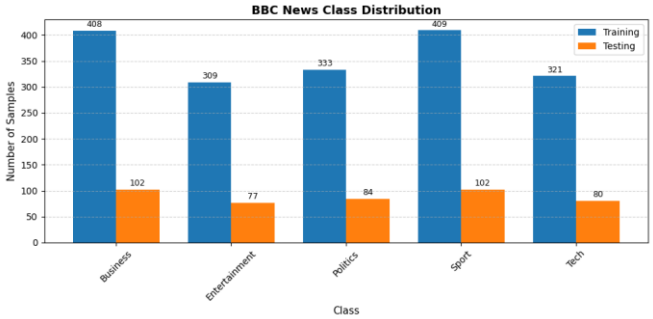


Fig. 3. BBC News class distribution

Since no official training-testing split was provided, a stratified 80/20 train-test split was employed. As shown in Fig. 3, the dataset shows mild class imbalance, reflecting a more realistic text classification scenario. The full dataset was utilized without subsampling.

3) DBpedia Ontology classification dataset: The DBpedia Ontology classification dataset is a large-scale dataset derived from structured content extracted from Wikipedia articles. Each document belongs to one of 14 ontology classes.

Each instance consists of a title and a content field, which were concatenated to form the textual input. The official training-testing split was preserved, and a class-balanced subset of approximately 10,000 samples was constructed by uniformly sampling from each class.

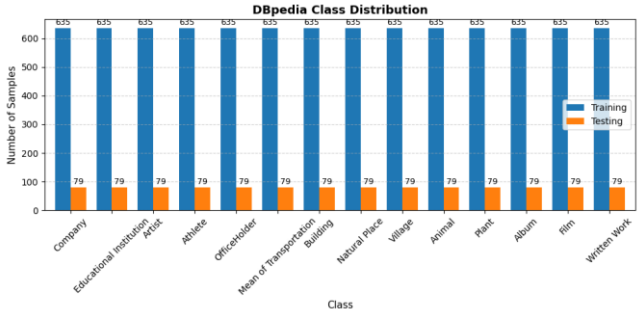


Fig. 4. DBpedia Ontology class distribution

As shown in Fig. 4, all ontology classes contain an equal number of samples. Compared to AG News and BBC News, the DBpedia Ontology classification dataset represents a more challenging task due to its larger number of classes and longer textual content.

C. Text Preprocessing

1) SVM and Distance-Based SVM: Before feature extraction, all datasets were preprocessed by using a standard preprocessing pipeline, as shown in Fig. 5, with the Natural Language Toolkit (NLTK) library. This pipeline reduces noise, standardizes textual representation, and improves feature quality for further vectorization and SVM classification.

The pipeline includes tokenization, normalization, noise removal, stopword elimination, and lemmatization to standardize textual input before feature extraction and classification. The preprocessing step includes:

- Raw Text Dataset: Input textual content from the original dataset.
- Tokenization: Split raw text into word-level tokens.
- Lowercasing: Convert all tokens to lower case to avoid duplicate feature representation.
- Remove non-alphabetic: remove punctuation, numbers, and symbols to reduce irrelevant features.
- Stopword Removal: Remove common and semantically uninformative features using NLTK's stopword list.

- Lemmatization: Reduce tokens to their base forms using WordNet lemmatizer, minimizing vocabulary size while preserving the semantic information.
- Cleaned Tokens: Rejoin processed tokens into a whitespace-separated string as input for the further text vectorization step.

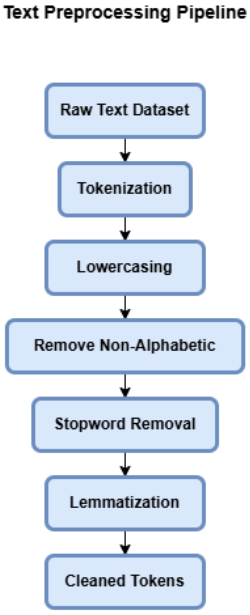


Fig. 5. Text preprocessing pipeline applied to all datasets.

2) LSTM: The raw text was converted into numerical representations by using a tokenizer. A word-level tokenizer was applied with a vocabulary size of 20,000 to ensure only the most frequent words are retained. The words that were not included in the vocabulary were mapped to an out-of-vocabulary (OOV) token.

The transformed sequences were then padded or truncated to a fixed maximum length of 500 tokens to ensure uniform length for the model input. The output of this step is a padded sequence of integer indices, which served as the input for the LSTM for the subsequent step.

D. Feature Representation

1) SVM and Distance-Based SVM: To enable the application of a machine learning algorithm, the pre-processed textual document must be transformed into a numerical representation, as a machine learning algorithm cannot directly consume raw text as input. The feature extraction was conducted by using TF-IDF, which is widely utilized in traditional text classification tasks.

TF-IDF is a statistical method that assigns weights to each word based on two components: term frequency (TF) and inverse document frequency (IDF). Term frequency determines the frequency of a term occurring in a document, while inverse document frequency penalizes the term that occurs frequently across many documents. As a result, the words that appear frequently in a specific document but rarely appear in the whole corpus will be assigned a higher weight to indicate their importance, while the common but less informative words will be assigned a lower score. This mechanism allows TF-IDF to highlight the discriminative terms that contribute more to distinguishing between classes.

Although text representation techniques, such as Word2Vec, GloVe, and contextual embeddings, such as BERT, have become more popular in NLP, TF-IDF was intentionally utilized to ensure the methodology consistency and controlled evaluation. The primary objective is to investigate the impact of different distance measures in the distance-based SVM framework. Introducing other methods like word embeddings or contextual embeddings will introduce additional complexity and confounding factors, making it more difficult to isolate the effect of distance measures.

Furthermore, TF-IDF generated high-sparse vectors, a numerical representation that well aligns with the geometric properties of SVM. This characteristic enables a clearer observation of how different distance metrics capture geometrical relationships in feature space. To ensure a fair and unbiased comparison, the TF-IDF vocabulary was learned only from the training data to prevent information leakage. The same vocabulary and weighting were then applied to transform the held-out test data. The TF-IDF was configured to limit the feature space to the 5000 most frequent terms. These settings were to balance the computational efficiency and representation capability. A moderate vocabulary size was selected to reduce sparsity while preserving sufficient important information for text classification. Since the primary objective is to evaluate the impact of different distance measures, the same TF-IDF configurations were applied in all experiments.

2) LSTM: The embedding layer was employed as the feature representation for LSTM. The embedding layer transforms the input word index into a dense 128-dimensional vector. This layer acts as a learnable feature representation module, where the words with similar semantic meaning are mapped closely in the embedding space, thus enabling the model to capture the semantic relationships between the words.

E. Model

1) Conventional SVM: The conventional SVM was employed as a baseline classifier and used as a reference for comparison with the proposed distance-based SVM variants for text classification. SVM is a supervised learning algorithm widely used for classification, known for its strong generalization and robustness to high-dimensional data. The workflow of conventional SVM was shown in Fig. 6.

The pre-processed and vectorized data served as the input for the SVM classifier. The core principle of SVM is to find an optimal separating hyperplane that maximizes the margin between different classes. The margin was defined as the distance between the hyperplane and the closest points from each class. Their closest points, known as support vectors, play a critical role in defining the decision. The support vectors represent the most informative points and are used as the

reference points in the later distance-based classification stage. This property is closely related to the margin maximization of SVM, where the support vectors were the closest point to the optimal separating hyperplane that maximize the margin between classes. Therefore, the discriminative information obtained from margin maximization was preserved when the support vectors are subsequently employed in the distance-based classification stage.

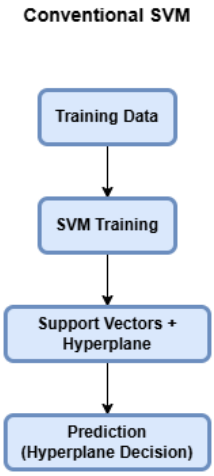


Fig. 6. Conventional SVM workflow

To enable a fair comparison of different types of data distributions, four commonly used kernel functions were employed: linear, polynomial, radial basis function (RBF), and sigmoid. The kernel function allows SVM to implicitly transform the data into a higher-dimensional feature space, enabling the data separation that the data are not linearly separable in the original space. The choice of kernel significantly affects the model's ability to capture the complex data patterns.

To ensure an unbiased evaluation, the regularization parameter was fixed in a wide range: C∈{1,10,100,1000,10000,100000}. The regularization parameter C allows SVM to control the trade-off between maximizing the margin and minimizing the classification error, thereby affecting model generalization and overfitting. The support vectors identified in the conventional SVM phase were retained and utilized in the distance-based SVM framework.

2) Distance-Based SVM: Fig. 7 demonstrates the overall workflow of the distance-based SVM. This approach shares the same training phase as conventional SVM for identifying support vectors. However, the distance-based SVM replaces the hyperplane-based decision with a distance-based mechanism. This modification does not modify the SVM optimization process but only alters the classification rule in the classification phase.

The distance-based SVM was a modified classification framework from the conventional SVM. Instead of relying on the hyperplane for classification prediction, this approach performs classification by computing the distance between the test instance and the support vectors for each class. This approach was inspired by Euclidean-SVM [6]. This study extends it by incorporating various distances instead of only the Euclidean distance.

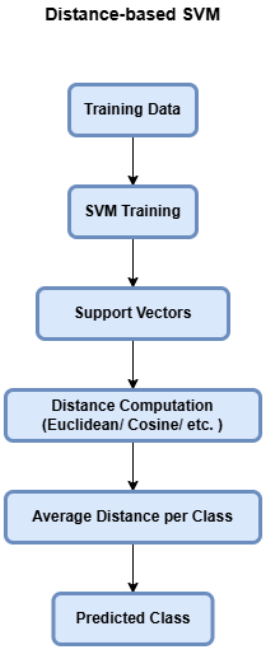


Fig. 7. Distance-based SVM workflow

In the proposed framework, the conventional SVM was first trained using the selected kernel and regularization parameter to identify the support vectors with their class labels. This training follows the standard SVM procedure and is identical to the conventional model. The aim of this training phase is solely to identify the support vectors, which are the most informative data near the decision boundary. During the classification phase, the hyperplane obtained from the training phase was not used for the classification prediction. Given a new unseen data x_{new}, the distance between the x_{new} and each support vector was computed. Subsequently, for each class y, the average distance between x_{new} and all the support vectors that belong to the class was calculated. Lastly, the predicted class, $\hat{y}$, is determined by selecting the class with the minimum average distance, as shown in Eq. (1):

$$\hat{y} = arg\ \min_{y}\left(\frac{1}{|SV_y|}\sum_{s\in SV_y} d(x_{new}, s)\right) \quad (1)$$

From Eq. (1), $|SV_y|$ denotes the number of the support vectors of class y, SV_y represents the set of support vectors of class y, while $d(x_{new}, s)$ indicates that the distance between the new instance x_{new} and the particular support vector. This decision rule assigns the new unseen data to the class that is closest to the support vectors on average. This strategy allows the classifier to decide based on the similarity relationships rather than a fixed hyperplane, which can be beneficial when the class distribution is complex and non-linear.

Unlike the Euclidean-SVM, which depends solely on the Euclidean distance, this study investigates the impact of different distance measures in the same framework. Various distance functions, including Euclidean distance, Manhattan distance, Chebyshev distance, Cosine distance, and Minkowski

distance, were investigated. Each distance defines a classifier and uses the same set of support vectors obtained from the training phase. To ensure a fair and unbiased evaluation with the conventional SVM, the distance-based SVM utilized identical kernel functions and regularization parameter configurations. All models are evaluated on a held-out test set.

3) LSTM Model: In addition to conventional SVM and distance-based SVM, a Long Short-Term Memory was employed as the deep learning baseline for text classification. LSTM is a special type of recurrent neural network that is designed to learn long-term dependencies in sequential data, making it suitable for text classification. The overall workflow of the LSTM was illustrated in Fig. 8.

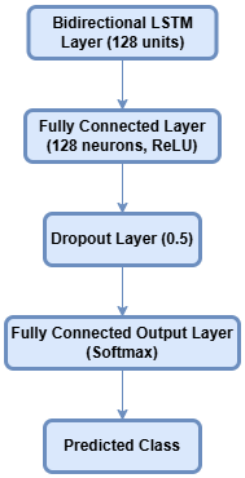


Fig. 8. LSTM workflow

The embedded sequences were passed to a bidirectional LSTM layer with 128 units. This layer processes the input sequence in both forward and backward directions, making it better at capturing the contextual information from both past and future words in the sequences. The output is then passed to a fully connected layer with 128 neurons and a ReLU activation function. This layer refines the learning representations for classification and learns the high-level features. A dropout layer with a dropout rate of 0.5 was then applied. This dropout layer helps to prevent overfitting by randomly deactivating a part of a neuron during training.

The model utilized a fully connected layer with a softmax activation function as the final layer. This layer produces a probability distribution over the target classes for multi-class classification. The model was trained by using the Adam optimizer with a learning rate of 0.001. The loss function utilized was categorical cross-entropy. Early stopping with a patience of 3 was applied to prevent overfitting.

F. Evaluation Metrics

The classification performance of the conventional SVM and distance-based SVM was evaluated by using two complementary metrics: accuracy and macro F1 score. Accuracy provides a general overview of model robustness, which measures the proportion of correctly classified instances to the total number of instances [see Eq. (2)]:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (2)$$

Macro F1 score captures the harmonic mean of precision and recall. For multi-class classification, the macro F1 score calculates the mean of the F1 score across all the classes independently, then averages them, treating all the classes evenly regardless of the class size [see Eq. (3)]:

$$F1_{macro} = \frac{F1_1 + F1_2 + \cdots + F1_N}{N} \quad (3)$$

where, N is the number of classes.

Accuracy was used as the primary metric to assess the performance comparison, while the macro F1 score provided additional insights for model robustness. Both metrics were computed on the test set for each kernel and regularization parameter configuration.

G. Experimental Setup and Evaluation Protocol

This section describes the experimental configuration used to evaluate the conventional SVM and distance-based SVM under a consistent and reproducible environment. The experiment was conducted in Python 3.13.5 by using scikit-learn 1.7.1, Numpy 2.3.2, pandas 2.3.1, and NLTK 3.9.1. To ensure reproducibility, a fixed random seed (random seed=42) was applied consistently across dataset splitting and SVM training. To ensure a fair comparison between SVM, distance-based SVM, and the LSTM model, the training and testing set was consistency between all models. The training data was used to build the model, while the held-out test set was utilized to evaluate the model's performance. To enable a fair comparison between SVM and distance-based SVM, both models:

- Identical TF-IDF numerical representation.
- Shared the same kernel function (linear, polynomial, RBF, and sigmoid) and regularization parameter ($C = 1, 10, 10^2, 10^3, 10^4, 10^5$).

To ensure a fair comparison between the SVM-based methods and LSTM, LSTM was conducted under the same dataset split and fixed random seed of 42. Final performance was evaluated on the test set by using accuracy and macro F1 score as the primary metrics.

III. Results and Discussion

A. Overall Performance Comparison Across SVM and Distance-Based SVM

This section presents the classification performance comparison of conventional SVM with each distance-based SVM variant, including Euclidean, Manhattan, Chebyshev, Cosine, and Minkowski across three text classification datasets: AG News, BBC News, and DBpedia Ontology classification dataset under various configurations. The performance was evaluated by using accuracy and F1 score, as presented in the Appendix (see Table I to Table VI) and illustrated in Fig. 9 to Fig. 14.

Fig. 9 to Fig. 11 demonstrate the accuracy of conventional SVM and distance-based SVM variants across different kernel and regularization parameter values on AG News, BBC News,

and DBpedia Ontology classification datasets. Each subfigure corresponds to a specific method, including conventional SVM and each distance-based variant. The individual curve represents a different kernel function. The horizontal axis shows the regularization parameter on a logarithmic scale, while the vertical axis demonstrates the accuracy in percentage. The exact numerical result is provided in the Appendix (see Table I to Table III).

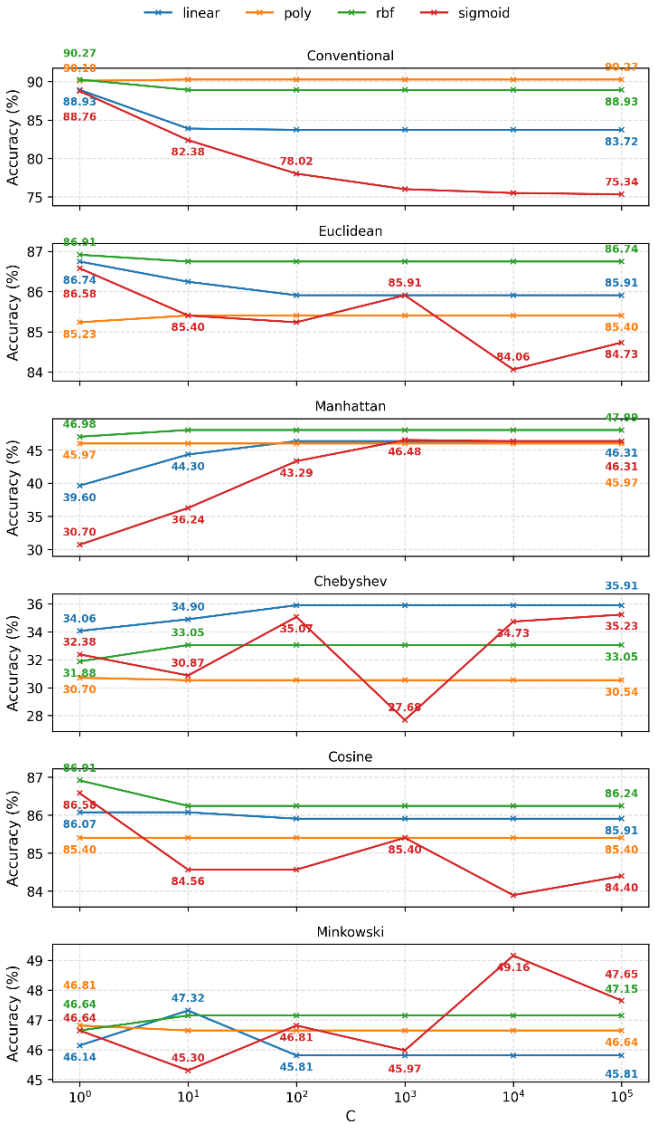


Fig. 9. Accuracy of the AG News dataset.

In addition to accuracy, the F1-score was also employed to account for class imbalance and provide a more comprehensive comparison of classification performance. Fig. 12 to Fig. 14 demonstrate the F1 score under an identical experimental configuration for all datasets. The exact numerical result is provided in the Appendix (see Table IV to Table VI).

Across all three datasets, the accuracy and the F1 score behaved very similarly under identical hyperparameter configurations within the same dataset. This indicates that the class imbalance did not distort the conclusion. The performance differences are not caused by the class imbalance, and confirms the robustness of the observation beyond accuracy alone.

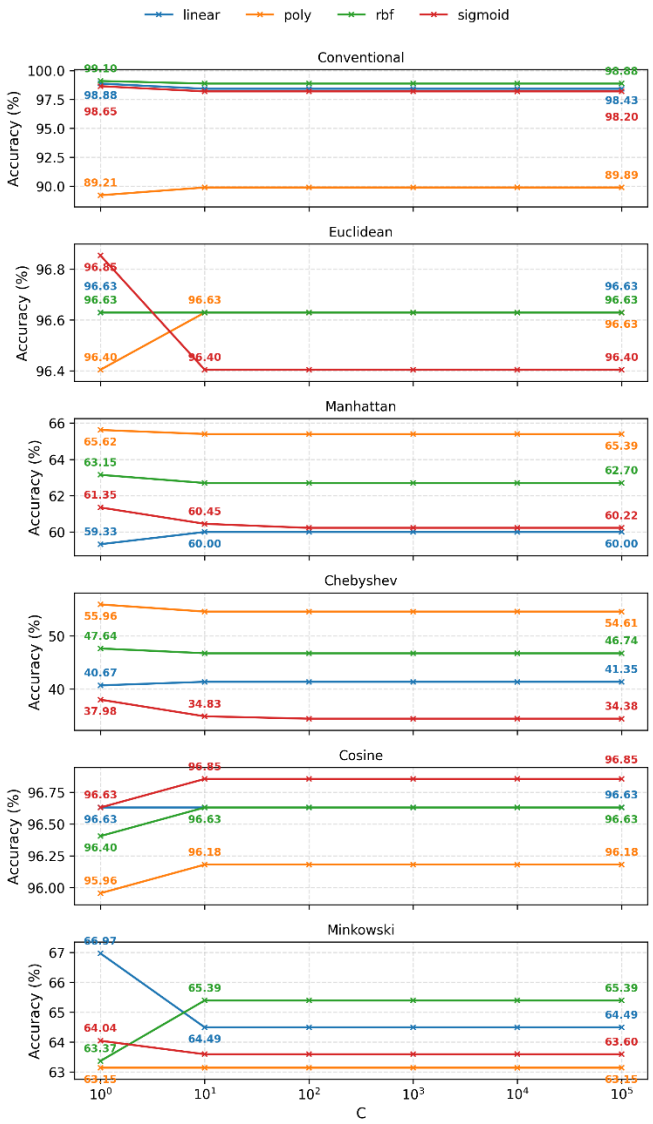


Fig. 10. Accuracy of the BBC News dataset.

Across all datasets, conventional SVM consistently achieves the peak in accuracy and in the F1 score, provided that the kernel and regularization parameters were carefully tuned. Specifically, the highest accuracy of 90.27% was obtained on the AG News dataset under the RBF kernel at C=1 and poly kernel at higher C values, 99.10% from BBC News by using the RBF kernel at C=1, and 95.48% on the DBpedia dataset under the linear kernel at C=10. However, the optimal kernel-regularization parameter combination varied across different datasets, showing its dataset-specific sensitivity.

Among the distance-based SVM variants, Euclidean and Cosine distance achieve competitive performance, closely approaching conventional SVM with only a little degradation in accuracy and F1 score across all datasets. Compared to the peak accuracy of conventional SVM, the performance

reduction remains bounded, with gaps of approximately 3-4% on the AG News dataset, 1-2% on the BBC News dataset, and 4-5% on the DBpedia dataset, indicating their strong discriminative capability. Their results remain consistently high, indicating their robustness in the text classification task.

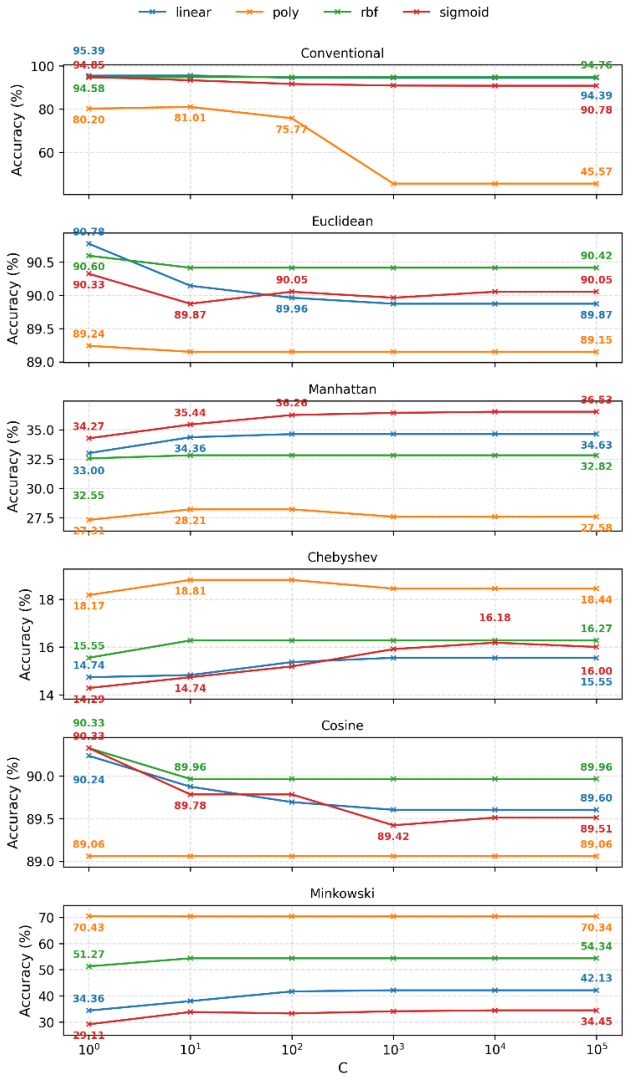


Fig. 11. Accuracy of the DBpedia Ontology classification dataset.

In contrast, the other distance-based SVM variants, including Manhattan, Chebyshev, and Minkowski distance-based SVM variants, consistently exhibit substantially lower performance across all datasets. These distances have low discriminative power in high-dimensional text representations, resulting in substantially lower accuracy and F1 score across kernel and regularization parameters.

Overall, the results indicate that the conventional SVM retains its position as the top performer when it is carefully tuned. The distance-based SVM, which uses Euclidean and Cosine distance, offers a strong trade-off between performance and robustness, making it suitable when excessive hyperparameter tuning is impractical and costly. This consistent trend was observed across all datasets with varying sizes and characteristics, indicating that the observed behaviors are not dataset-specific.

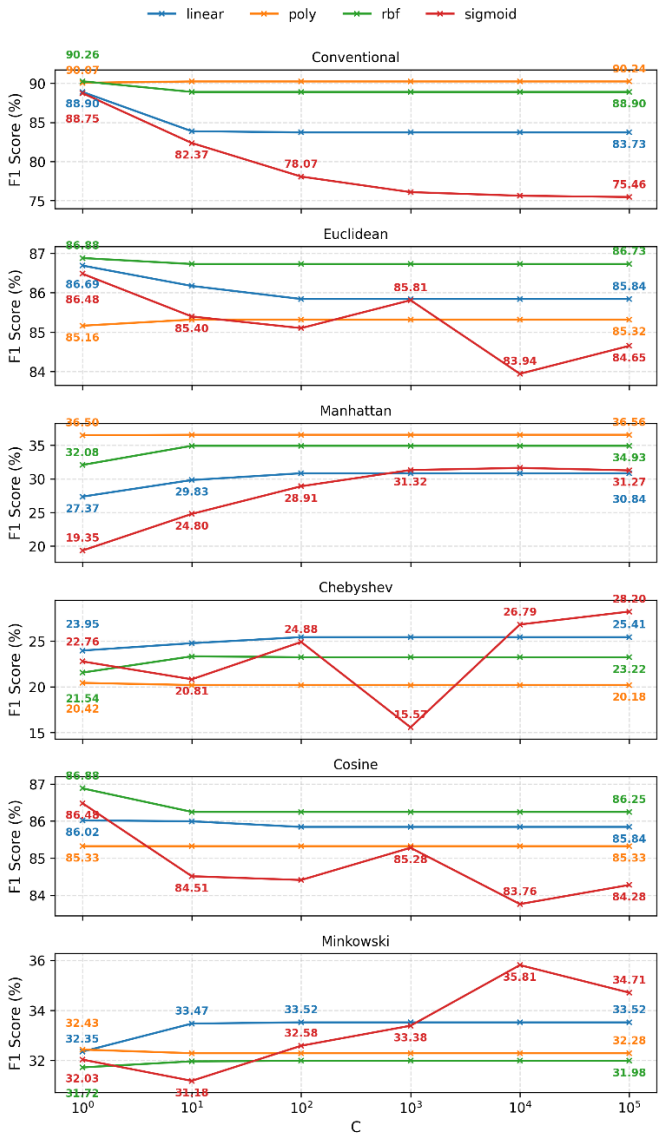


Fig. 12. F1 score of the AG News dataset.

B. Sensitivity of Conventional SVM to Kernel Functions and Regularization Parameters

This section evaluates the classification performance with respect to different kernel functions and regularization parameters, emphasizing the performance stability rather than the peak performance. The sensitivity of performance to hyperparameters is a critical issue for SVM classification tasks, as an inappropriate kernel function or regularization parameter may lead to substantial degradation in performance.

The conventional SVM exhibits a noticeable sensitivity to both kernel function and regularization parameter across all three datasets. While high accuracy and F1 score were obtained when the optimal combination of kernel and regularization parameter was chosen, the performance fluctuates significantly when the suboptimal combination of hyperparameters was

employed. For instance, on the AG News dataset, the sigmoid kernel demonstrates a noticeable performance degradation, dropping from 88.76% to 75.34% as the value of the regularization parameter increases. Similarly, for the BBC News dataset, the linear, RBF, and sigmoid kernels exhibit a near-optimal performance with accuracy approximately at 98-99%, but a clear performance gap is observed when switching the kernel to polynomial, where accuracy is maintained at 89-90%. In addition, on the DBpedia ontology classification dataset, the polynomial kernel demonstrates a severe decline in performance, dropping from 80.20% to 45.57% as the value of the regularization parameter increases. These observations highlight that conventional SVM performance was heavily dependent on dataset-specific hyperparameter tuning.

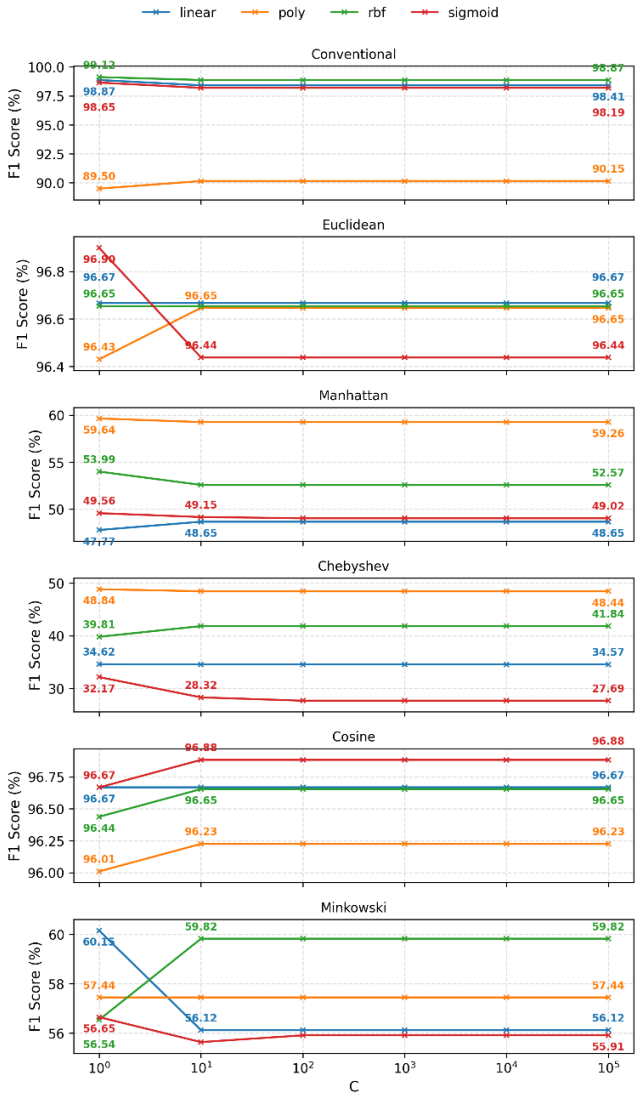


Fig. 13. F1 score of the BBC News dataset.

In contrast, the distance-based SVM variants, which utilize Euclidean and Cosine distance, demonstrate a substantially reduced fluctuation in performance across different kernels and regularization strengths. As illustrated in Fig. 9 to Fig. 14, their accuracy and F1 score exhibit nearly flat performance curves over all kernels and regularization strengths across all datasets. This minimal performance variation was bounded within a narrow margin, generally within 0-3%, suggesting that these methods exhibit more stable behavior under hyperparameter changing. This robustness is particularly evident in the DBpedia dataset, where Euclidean and Cosine exhibit a stable performance, even though the conventional SVM experiences a sharp decline in performance.

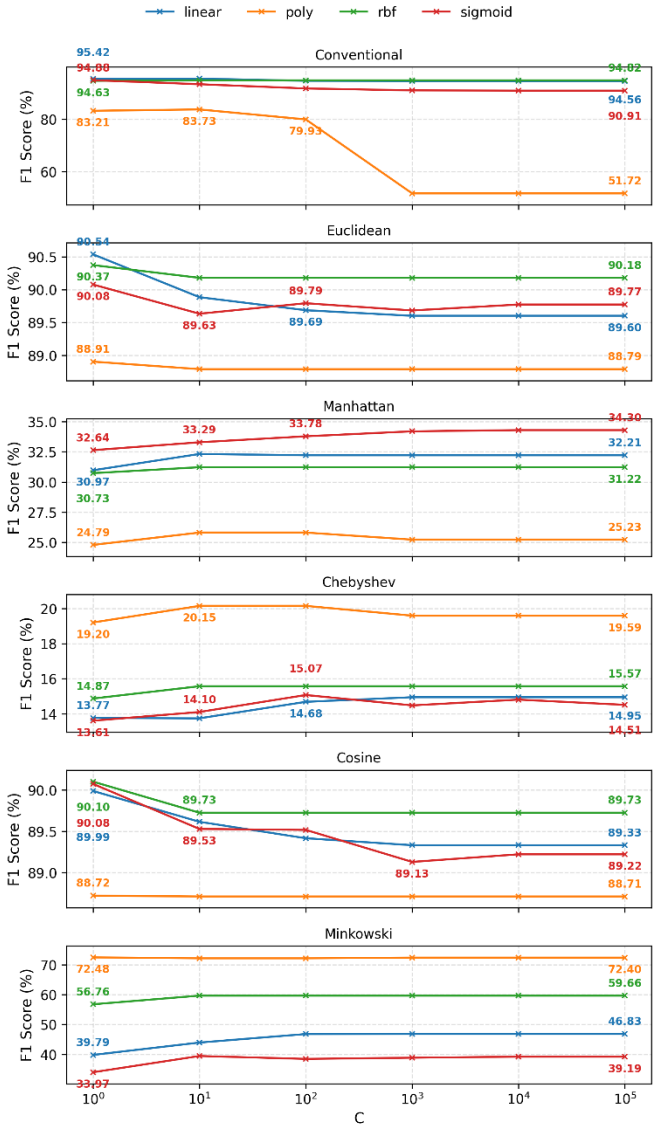


Fig. 14. F1 score of the DBpedia Ontology classification dataset.

The other distance-based SVM variants, which employed Manhattan, Chebyshev, and Minkowski distance, demonstrate low performance regardless of kernel and regularization parameter. The consistency across the different hyperparameter configurations suggests that for these distances, the choice of distance measures itself essentially limits the discriminative power. These distances may be less sensitive to the geometry in the high-dimensional, sparse TF-IDF vectors, making the average distance to support vectors not vary enough to support accurate classification. As a result, adjusting the kernel function and regularization parameter does not bring a significant impact on improving the classification performance.

Overall, these results clearly indicate the trade-off between peak performance and performance robustness. The conventional SVM can achieve peak performance when hyperparameter tuning is optimal. In contrast, Euclidean and Cosine-based SVM variants offer a more stable alternative, exhibiting lower variation across different kernel functions and regularization parameters, making them suitable for environments where exhaustive hyperparameter tuning was impractical.

C. Interpretation of Distance Metrics in High-Dimensional Text Classification

The text document was represented as a TF-IDF numerical representation. These representations produce high-dimensional and sparse vectors, where most of the values are zero. In this case, the choice of distance metric plays an important role in classification performance.

The Euclidean and Cosine distance-based SVM variants demonstrate strong and consistent performance across all datasets. Cosine distance measures the similarity of documents based on the orientation of feature vectors, emphasizing the relative term utilized pattern, making it effective in capturing the semantic similarity. This property plays an important role in text classification, as documents in the same category share a similar term distribution. Euclidean distance also performs well due to TF-IDF representation, which places the documents on a similar scale, making the distance comparison more meaningful. The document with similar patterns stays closer together, while the document from a different class is farther, leading to a reliable decision boundary for text classification.

Furthermore, distance-based SVMs may be more suitable for high-dimensional sparse data, given document representations such as TF-IDF, which primarily capture word-occurrence patterns rather than dense numerical relationships. In such feature spaces, the documents from the same classes often form local clusters with similar feature distributions. Although the performance is slightly lower than the peak accuracy of fine-tuned conventional SVM, their reduced sensitivity to kernel function and regularization parameters makes them more suitable in scenarios where extensive hyperparameter tuning or computational resources are limited.

In contrast, the distance-based SVM variants that employ Manhattan, Chebyshev, and Minkowski distance consistently perform weakly across all datasets. This behavior may relate to the geometric properties of high-dimensional sparse TF-IDF representations. Manhattan distance accumulates the absolute differences across all dimensions, making many small feature differences contribute equally to the final distance. Chebyshev distance only focuses on the feature with maximum distance, which means that the classification is dominated by one word, thus ignoring the overall document similarity, which significantly degrades the classification performance since the document's meaning comes from many words, not one. Similarly, the Minkowski distance with p=3 places greater emphasis on larger feature differences, which may distort the similarity relationships among sparse document vectors. As a result, these distances may be less effective in preserving meaningful document relationships in sparse text representations, regardless of kernel choice or regularization strength. This indicates that for distance-based SVM, the choice of distance plays a more important role than the kernel and regularization in high-dimensional classification datasets.

For the text classification task, if the environment is feasible for exhaustive hyperparameter tuning, the conventional SVM is recommended. However, when a rapid development is prioritized or when resources are limited and costly, the Euclidean and Cosine distance SVM variants are a preferable alternative. For example, it is suitable for the large-scale text system and resource-constrained environments.

D. Performance Comparison with LSTM

To further evaluate the effectiveness of the distance-based SVM variants, a comparison was conducted with a deep learning method, namely Long Short-Term Memory (LSTM).

Unlike the SVM-based methods, which were conducted under varied hyperparameter configurations, the LSTM model was trained using stochastic gradient-based optimization. As a result, the LSTM was reported with the accuracy and F1 Score under the same random seed as SVM, while the SVM-based methods were reported with their best performing configurations under varying hyperparameter scenarios. The results were reported in Fig. 15 and Fig. 16.

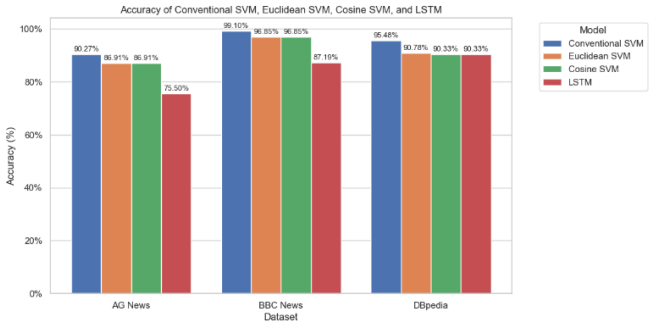


Fig. 15. Accuracy of Conventional SVM, Euclidean SVM, Cosine SVM, and LSTM.

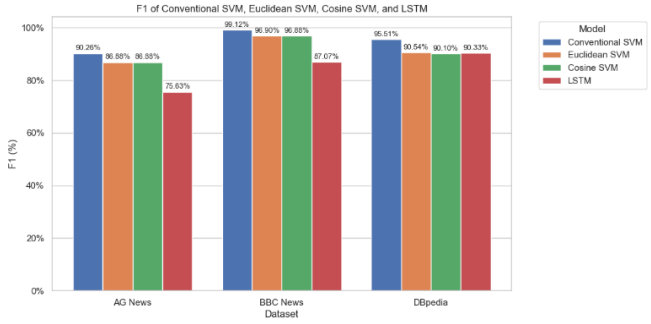


Fig. 16. F1 score of Conventional SVM, Euclidean SVM, Cosine SVM, and LSTM.

Across all the datasets, the SVM-based approaches consistently outperform LSTM. For the AG News dataset, the conventional SVM achieves an accuracy of 90.27%, and the F1 score of 90.26%; the Euclidean and Cosine-based SVMs achieve slightly lower but comparable performance with an accuracy of 86.91% and the F1 score of 86.88%. In contrast, the LSTM model achieves a notably lower accuracy of 75.50% and the F1 Score of 75.63%.

A similar trend was observed in the BBC News dataset. The conventional SVM achieves a near-perfect performance with an accuracy of 99.10% and the F1 score of 99.12%. The Euclidean and Cosine-based SVMs achieved a comparable performance with a slight degradation, with accuracy and F1 score falling between 96.8% and 97%. The LSTM model only achieves an accuracy of 87.19% and an F1 score of 87.07%, which is much lower than the SVM-based methods, indicating it is less effective in capturing the patterns compared to SVM-based methods.

For the DBpedia dataset, although the LSTM exhibits an obvious improvement compared to other datasets with an accuracy and F1 score of 90.33%, it still falls behind compared to SVM models with accuracy and F1 score above 95%, and comparable with the distance-based SVM, which includes Euclidean and Cosine that achieve similar accuracy and F1 Score.

Overall, the results demonstrate that the SVM-based models, including both conventional and distance-based variants, including Euclidean and Cosine, outperform the LSTM model under these three datasets. Notably, the Euclidean and Cosine-based SVM variants provide a strong balance between performance and stability, achieving results close to conventional SVM while reducing the sensitivity to hyperparameter changes.

IV. Conclusion

This study evaluates both conventional SVM and distance-based SVM, which incorporate various distance measures on three text classification datasets with varying size, class complexity, and text length. The aim is to assess both the peak performance and robustness to different kernel functions and regularization parameters.

From the experimental results observed across three text classification datasets, conventional SVM can achieve the highest peak accuracy provided the kernel functions and regularization parameters were fine-tuned. However, the performance exhibits substantial degradation when suboptimal hyperparameters are chosen. In contrast, the distance-based SVM variants that employed Euclidean and Cosine distance achieve a relatively slightly lower peak accuracy but maintain consistent performance across various kernels and regularization parameters. Other distance-based SVM variants, including Manhattan, Chebyshev, and Minkowski distance-based SVM, consistently performed poorly across all datasets.

The observed robustness of Euclidean and Cosine-based SVM variants reduces the sensitivity to kernel function and regularization parameter changes, making them suitable for scenarios where excessive hyperparameter tuning is computationally costly or impractical. This study provides practical guidance for selecting the SVM variants for text classification tasks, especially for resource-restricted environments.

In addition, the comparison with LSTM models shows that SVM-based models, including the conventional SVM and the high-performing distance-based variants (Euclidean and Cosine), consistently outperform LSTM models in both accuracy and F1 score across all datasets. These findings suggest that the traditional SVM-based models remain highly effective for the text classification task under the studied settings.

Future work can be extended by conducting a comprehensive evaluation with a deep learning approach, such as an advanced recurrent network or transformer-based models (e.g., BERT). The evaluation should be conducted under a unified experimental protocol, including consistent data preprocessing, train-test splits, and evaluation metrics to ensure a fair and systematic comparison across different models. In addition, further investigation can be carried out into other text representation techniques and large-scale text datasets to assess the generalization capability of the proposed approaches.

In short, the Euclidean- and Cosine-based SVM variants offer a robust and reliable alternative to conventional SVMs for text classification applications where hyperparameter tuning is limited.

References

[1] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” Lect. Notes Comput. Sci., vol. 1398, pp. 137–142, 1998.

[2] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.

[3] Y. Wahba, N. Madhavji, and J. Steinbacher, “A Comparison of SVM Against Pre-trained Language Models (PLMs) for Text Classification Tasks,” Lect. Notes Comput. Sci., vol. 13811 LNCS, pp. 304–313, 2023.

[4] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, “Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review,” Algorithms, vol. 16, no. 5, pp. 1–28, 2023.

[5] P. Rebentrost, M. Mohseni, and S. Lloyd, “Quantum support vector machine for big data classification,” Phys. Rev. Lett., vol. 113, no. 3, pp. 1–5, 2014.

[6] L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, “An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization,” Appl. Intell., vol. 37, no. 1, pp. 80–99, 2012.

[7] L. H. Lee, R. Rajkumar, L. H. Lo, C. H. Wan, and D. Isa, “Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach,” Expert Syst. Appl., vol. 40, no. 6, pp. 1925–1934, 2013.

[8] N. Amaya-Tejera, M. Gamarra, J. I. Vélez, and E. Zurek, “A distance-based kernel for classification via Support Vector Machines,” Front. Artif. Intell., vol. 7, 2024.

[9] W. Liu, S. Liang, and X. Qin, “Weighted p-norm distance t kernel SVM classification algorithm based on improved polarization,” Sci. Rep., vol. 12, no. 1, pp. 1–16, 2022.

[10] X. Zhang, J. Zhao, and Y. LeCun, “Character-level Convolutional Networks for Text Classification,” in Proceedings of the 29th International Conference on Neural Information Processing Systems, 2016, vol. 1, pp. 649–657.

[11] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” ACM Int. Conf. Proceeding Ser., vol. 148, pp. 377–384, 2006.

APPENDIX

TABLE I. ACCURACY OF AG NEWS DATASET

Method	Kernel	C=1	C=10	C=100	C=1000	C=10000	C=100000
Conventional	linear	88.93%	83.89%	83.72%	83.72%	83.72%	83.72%
Conventional	poly	90.10%	90.27%	90.27%	90.27%	90.27%	90.27%
Conventional	rbf	90.27%	88.93%	88.93%	88.93%	88.93%	88.93%
Conventional	sigmoid	88.76%	82.38%	78.02%	76.01%	75.50%	75.34%
Euclidean	linear	86.74%	86.24%	85.91%	85.91%	85.91%	85.91%
Euclidean	poly	85.23%	85.40%	85.40%	85.40%	85.40%	85.40%
Euclidean	rbf	86.91%	86.74%	86.74%	86.74%	86.74%	86.74%
Euclidean	sigmoid	86.58%	85.40%	85.23%	85.91%	84.06%	84.73%
Manhattan	linear	39.60%	44.30%	46.31%	46.31%	46.31%	46.31%
Manhattan	poly	45.97%	45.97%	45.97%	45.97%	45.97%	45.97%
Manhattan	rbf	46.98%	47.99%	47.99%	47.99%	47.99%	47.99%
Manhattan	sigmoid	30.70%	36.24%	43.29%	46.48%	46.31%	46.31%
Chebyshev	linear	34.06%	34.90%	35.91%	35.91%	35.91%	35.91%
Chebyshev	poly	30.70%	30.54%	30.54%	30.54%	30.54%	30.54%
Chebyshev	rbf	31.88%	33.05%	33.05%	33.05%	33.05%	33.05%
Chebyshev	sigmoid	32.38%	30.87%	35.07%	27.68%	34.73%	35.23%
Cosine	linear	86.07%	86.07%	85.91%	85.91%	85.91%	85.91%
Cosine	poly	85.40%	85.40%	85.40%	85.40%	85.40%	85.40%
Cosine	rbf	86.91%	86.24%	86.24%	86.24%	86.24%	86.24%
Cosine	sigmoid	86.58%	84.56%	84.56%	85.40%	83.89%	84.40%
Minkowski	linear	46.14%	47.32%	45.81%	45.81%	45.81%	45.81%
Minkowski	poly	46.81%	46.64%	46.64%	46.64%	46.64%	46.64%
Minkowski	rbf	46.64%	47.15%	47.15%	47.15%	47.15%	47.15%
Minkowski	sigmoid	46.64%	45.30%	46.81%	45.97%	49.16%	47.65%

TABLE II. ACCURACY OF BBC NEWS DATASET

Method	Kernel	C=1	C=10	C=100	C=1000	C=10000	C=100000
Conventional	linear	98.88%	98.43%	98.43%	98.43%	98.43%	98.43%
Conventional	poly	89.21%	89.89%	89.89%	89.89%	89.89%	89.89%
Conventional	rbf	99.10%	98.88%	98.88%	98.88%	98.88%	98.88%
Conventional	sigmoid	98.65%	98.20%	98.20%	98.20%	98.20%	98.20%
Euclidean	linear	96.63%	96.63%	96.63%	96.63%	96.63%	96.63%
Euclidean	poly	96.40%	96.63%	96.63%	96.63%	96.63%	96.63%
Euclidean	rbf	96.63%	96.63%	96.63%	96.63%	96.63%	96.63%
Euclidean	sigmoid	96.85%	96.40%	96.40%	96.40%	96.40%	96.40%
Manhattan	linear	59.33%	60.00%	60.00%	60.00%	60.00%	60.00%
Manhattan	poly	65.62%	65.39%	65.39%	65.39%	65.39%	65.39%
Manhattan	rbf	63.15%	62.70%	62.70%	62.70%	62.70%	62.70%
Manhattan	sigmoid	61.35%	60.45%	60.22%	60.22%	60.22%	60.22%

Chebyshev	linear	40.67%	41.35%	41.35%	41.35%	41.35%	41.35%
Chebyshev	poly	55.96%	54.61%	54.61%	54.61%	54.61%	54.61%
Chebyshev	rbf	47.64%	46.74%	46.74%	46.74%	46.74%	46.74%
Chebyshev	sigmoid	37.98%	34.83%	34.38%	34.38%	34.38%	34.38%
Cosine	linear	96.63%	96.63%	96.63%	96.63%	96.63%	96.63%
Cosine	poly	95.96%	96.18%	96.18%	96.18%	96.18%	96.18%
Cosine	rbf	96.40%	96.63%	96.63%	96.63%	96.63%	96.63%
Cosine	sigmoid	96.63%	96.85%	96.85%	96.85%	96.85%	96.85%
Minkowski	linear	66.97%	64.49%	64.49%	64.49%	64.49%	64.49%
Minkowski	poly	63.15%	63.15%	63.15%	63.15%	63.15%	63.15%
Minkowski	rbf	63.37%	65.39%	65.39%	65.39%	65.39%	65.39%
Minkowski	sigmoid	64.04%	63.60%	63.60%	63.60%	63.60%	63.60%

TABLE III. ACCURACY OF DBPEDIA ONTOLOGY CLASSIFICATION DATASET

Method	Kernel	C=1	C=10	C=100	C=1000	C=10000	C=100000
Conventional	linear	95.39%	95.48%	94.48%	94.39%	94.39%	94.39%
Conventional	poly	80.20%	81.01%	75.77%	45.57%	45.57%	45.57%
Conventional	rbf	94.58%	94.76%	94.76%	94.76%	94.76%	94.76%
Conventional	sigmoid	94.85%	93.31%	91.59%	90.87%	90.78%	90.78%
Euclidean	linear	90.78%	90.14%	89.96%	89.87%	89.87%	89.87%
Euclidean	poly	89.24%	89.15%	89.15%	89.15%	89.15%	89.15%
Euclidean	rbf	90.60%	90.42%	90.42%	90.42%	90.42%	90.42%
Euclidean	sigmoid	90.33%	89.87%	90.05%	89.96%	90.05%	90.05%
Manhattan	linear	33.00%	34.36%	34.63%	34.63%	34.63%	34.63%
Manhattan	poly	27.31%	28.21%	28.21%	27.58%	27.58%	27.58%
Manhattan	rbf	32.55%	32.82%	32.82%	32.82%	32.82%	32.82%
Manhattan	sigmoid	34.27%	35.44%	36.26%	36.44%	36.53%	36.53%
Chebyshev	linear	14.74%	14.83%	15.37%	15.55%	15.55%	15.55%
Chebyshev	poly	18.17%	18.81%	18.81%	18.44%	18.44%	18.44%
Chebyshev	rbf	15.55%	16.27%	16.27%	16.27%	16.27%	16.27%
Chebyshev	sigmoid	14.29%	14.74%	15.19%	15.91%	16.18%	16.00%
Cosine	linear	90.24%	89.87%	89.69%	89.60%	89.60%	89.60%
Cosine	poly	89.06%	89.06%	89.06%	89.06%	89.06%	89.06%
Cosine	rbf	90.33%	89.96%	89.96%	89.96%	89.96%	89.96%
Cosine	sigmoid	90.33%	89.78%	89.78%	89.42%	89.51%	89.51%
Minkowski	linear	34.36%	37.97%	41.68%	42.13%	42.13%	42.13%
Minkowski	poly	70.43%	70.34%	70.34%	70.34%	70.34%	70.34%
Minkowski	rbf	51.27%	54.34%	54.34%	54.34%	54.34%	54.34%
Minkowski	sigmoid	29.11%	33.82%	33.27%	34.09%	34.45%	34.45%

TABLE IV. F1 SCORE OF AG NEWS DATASET

Method	Kernel	C=1	C=10	C=100	C=1000	C=10000	C=100000
Conventional	linear	88.90%	83.88%	83.73%	83.73%	83.73%	83.73%
Conventional	poly	90.07%	90.24%	90.24%	90.24%	90.24%	90.24%
Conventional	rbf	90.26%	88.90%	88.90%	88.90%	88.90%	88.90%
Conventional	sigmoid	88.75%	82.37%	78.07%	76.08%	75.63%	75.46%
Euclidean	linear	86.69%	86.17%	85.84%	85.84%	85.84%	85.84%
Euclidean	poly	85.16%	85.32%	85.32%	85.32%	85.32%	85.32%
Euclidean	rbf	86.88%	86.73%	86.73%	86.73%	86.73%	86.73%
Euclidean	sigmoid	86.48%	85.40%	85.10%	85.81%	83.94%	84.65%
Manhattan	linear	27.37%	29.83%	30.84%	30.84%	30.84%	30.84%
Manhattan	poly	36.50%	36.56%	36.56%	36.56%	36.56%	36.56%
Manhattan	rbf	32.08%	34.93%	34.93%	34.93%	34.93%	34.93%
Manhattan	sigmoid	19.35%	24.80%	28.91%	31.32%	31.64%	31.27%
Chebyshev	linear	23.95%	24.75%	25.41%	25.41%	25.41%	25.41%
Chebyshev	poly	20.42%	20.18%	20.18%	20.18%	20.18%	20.18%
Chebyshev	rbf	21.54%	23.32%	23.22%	23.22%	23.22%	23.22%
Chebyshev	sigmoid	22.76%	20.81%	24.88%	15.57%	26.79%	28.20%
Cosine	linear	86.02%	85.99%	85.84%	85.84%	85.84%	85.84%
Cosine	poly	85.33%	85.33%	85.33%	85.33%	85.33%	85.33%
Cosine	rbf	86.88%	86.25%	86.25%	86.25%	86.25%	86.25%
Cosine	sigmoid	86.48%	84.51%	84.41%	85.28%	83.76%	84.28%
Minkowski	linear	32.35%	33.47%	33.52%	33.52%	33.52%	33.52%
Minkowski	poly	32.43%	32.28%	32.28%	32.28%	32.28%	32.28%
Minkowski	rbf	31.72%	31.96%	31.98%	31.98%	31.98%	31.98%
Minkowski	sigmoid	32.03%	31.18%	32.58%	33.38%	35.81%	34.71%

TABLE V. F1 SCORE OF BBC NEWS DATASET

Method	Kernel	C=1	C=10	C=100	C=1000	C=10000	C=100000
Conventional	linear	98.87%	98.41%	98.41%	98.41%	98.41%	98.41%
Conventional	poly	89.50%	90.15%	90.15%	90.15%	90.15%	90.15%
Conventional	rbf	99.12%	98.87%	98.87%	98.87%	98.87%	98.87%
Conventional	sigmoid	98.65%	98.19%	98.19%	98.19%	98.19%	98.19%
Euclidean	linear	96.67%	96.67%	96.67%	96.67%	96.67%	96.67%
Euclidean	poly	96.43%	96.65%	96.65%	96.65%	96.65%	96.65%
Euclidean	rbf	96.65%	96.65%	96.65%	96.65%	96.65%	96.65%
Euclidean	sigmoid	96.90%	96.44%	96.44%	96.44%	96.44%	96.44%
Manhattan	linear	47.77%	48.65%	48.65%	48.65%	48.65%	48.65%
Manhattan	poly	59.64%	59.26%	59.26%	59.26%	59.26%	59.26%
Manhattan	rbf	53.99%	52.57%	52.57%	52.57%	52.57%	52.57%
Manhattan	sigmoid	49.56%	49.15%	49.02%	49.02%	49.02%	49.02%

Chebyshev	linear	34.62%	34.57%	34.57%	34.57%	34.57%	34.57%
Chebyshev	poly	48.84%	48.44%	48.44%	48.44%	48.44%	48.44%
Chebyshev	rbf	39.81%	41.84%	41.84%	41.84%	41.84%	41.84%
Chebyshev	sigmoid	32.17%	28.32%	27.69%	27.69%	27.69%	27.69%
Cosine	linear	96.67%	96.67%	96.67%	96.67%	96.67%	96.67%
Cosine	poly	96.01%	96.23%	96.23%	96.23%	96.23%	96.23%
Cosine	rbf	96.44%	96.65%	96.65%	96.65%	96.65%	96.65%
Cosine	sigmoid	96.67%	96.88%	96.88%	96.88%	96.88%	96.88%
Minkowski	linear	60.15%	56.12%	56.12%	56.12%	56.12%	56.12%
Minkowski	poly	57.44%	57.44%	57.44%	57.44%	57.44%	57.44%
Minkowski	rbf	56.54%	59.82%	59.82%	59.82%	59.82%	59.82%
Minkowski	sigmoid	56.65%	55.63%	55.91%	55.91%	55.91%	55.91%

TABLE VI. F1 SCORE OF DBPEDIA ONTOLOGY CLASSIFICATION DATASET

Method	**Kernel**	**C=1**	**C=10**	**C=100**	**C=1000**	**C=10000**	**C=100000**
Conventional	linear	95.42%	95.51%	94.63%	94.56%	94.56%	94.56%
Original	poly	83.21%	83.73%	79.93%	51.72%	51.72%	51.72%
Original	rbf	94.63%	94.82%	94.82%	94.82%	94.82%	94.82%
Original	sigmoid	94.88%	93.35%	91.73%	91.01%	90.91%	90.91%
Euclidean	linear	90.54%	89.89%	89.69%	89.60%	89.60%	89.60%
Euclidean	poly	88.91%	88.79%	88.79%	88.79%	88.79%	88.79%
Euclidean	rbf	90.37%	90.18%	90.18%	90.18%	90.18%	90.18%
Euclidean	sigmoid	90.08%	89.63%	89.79%	89.68%	89.77%	89.77%
Manhattan	linear	30.97%	32.32%	32.21%	32.21%	32.21%	32.21%
Manhattan	poly	24.79%	25.81%	25.81%	25.23%	25.23%	25.23%
Manhattan	rbf	30.73%	31.22%	31.22%	31.22%	31.22%	31.22%
Manhattan	sigmoid	32.64%	33.29%	33.78%	34.19%	34.30%	34.30%
Chebyshev	linear	13.77%	13.74%	14.68%	14.95%	14.95%	14.95%
Chebyshev	poly	19.20%	20.15%	20.15%	19.59%	19.59%	19.59%
Chebyshev	rbf	14.87%	15.57%	15.57%	15.57%	15.57%	15.57%
Chebyshev	sigmoid	13.61%	14.10%	15.07%	14.48%	14.81%	14.51%
Cosine	linear	89.99%	89.62%	89.42%	89.33%	89.33%	89.33%
Cosine	poly	88.72%	88.71%	88.71%	88.71%	88.71%	88.71%
Cosine	rbf	90.10%	89.73%	89.73%	89.73%	89.73%	89.73%
Cosine	sigmoid	90.08%	89.53%	89.52%	89.13%	89.22%	89.22%
Minkowski	linear	39.79%	43.93%	46.79%	46.83%	46.83%	46.83%
Minkowski	poly	72.48%	72.20%	72.20%	72.40%	72.40%	72.40%
Minkowski	rbf	56.76%	59.66%	59.66%	59.66%	59.66%	59.66%
Minkowski	sigmoid	33.97%	39.42%	38.45%	38.84%	39.19%	39.19%

Pareto-Optimized Model Predictive Control for Dynamically Feasible Three-Dimensional Trajectory Generation in Robotic Manipulators

Zeinel Momynkulov[1], Sayat Ibrayev[2], Azizah Suliman[3], Yegenberdi Tenizbayev[4], Batyrkhan Omarov[5]*
Joldasbekov Institute of Mechanics and Engineering, Kazakhstan[1, 2, 5]
International Information Technology University, Kazakhstan[1, 5]
Asia Metropolitan University, Malaysia[3]
Central Asian Innovative University, Kazakhstan[4]
Narxoz University, Kazakhstan[5]
Khalel Dosmukhamedov Atyrau University, Kazakhstan[5]

***Abstract*—This study presents a Pareto-optimized Model Predictive Control (MPC) framework for dynamically feasible three-dimensional trajectory generation in robotic manipulators operating under physical constraints. Unlike conventional interpolation-based methods that emphasize geometric smoothness while neglecting system dynamics, the proposed approach integrates a second-order discrete-time model with explicit constraints on position, velocity, and acceleration, ensuring physically consistent motion profiles. A multi-objective optimization strategy is introduced, combining grid search with Pareto front analysis to systematically tune key MPC parameters, including prediction horizon and discretization step. This enables a principled trade-off between tracking accuracy and control effort, addressing a critical limitation in existing MPC implementations that rely on heuristic parameter selection. Experimental results demonstrate that the proposed method achieves competitive tracking performance while significantly improving trajectory smoothness and reducing acceleration peaks compared to spline-based and linear interpolation approaches. The framework maintains real-time feasibility with computation times below 20 ms per control cycle, making it suitable for practical deployment in robotic systems. Furthermore, the integration of learning-based trajectory generation highlights the adaptability of the approach in complex and dynamic environments. Overall, the proposed methodology offers a scalable, interpretable, and computationally efficient solution that bridges the gap between geometric trajectory planning and physically realizable robotic motion, contributing to the advancement of control-aware trajectory generation in modern robotic applications.**

Keywords*—Model Predictive Control; trajectory planning; robotic manipulators; Pareto optimization; multi-objective optimization; dynamic constraints; real-time control; 3d motion planning

I. Introduction

Robotic manipulators have become indispensable components of modern intelligent systems, playing a pivotal role in industrial automation, precision manufacturing, medical robotics, and space exploration. Their increasing deployment in complex and dynamic environments has significantly raised the demands placed on motion planning algorithms, particularly in terms of accuracy, robustness, and physical feasibility. In such settings, trajectory generation is no longer a purely geometric problem but must incorporate the physical limitations of actuators and the dynamic behavior of the system [1]. Consequently, the development of advanced trajectory planning strategies that ensure both precision and feasibility has emerged as a critical research direction in robotics.

Traditional trajectory generation approaches, including polynomial interpolation, cubic splines, and B-splines, have been widely adopted due to their computational simplicity and ability to produce smooth paths. These methods are particularly effective in scenarios where geometric continuity is the primary concern. However, they inherently lack awareness of system dynamics and constraints, such as bounded velocity, acceleration, and actuator capabilities [2]. As a result, trajectories generated using purely geometric techniques may be mathematically smooth yet physically infeasible when executed on real robotic platforms. This limitation becomes especially pronounced in high-speed or safety-critical applications, where violations of dynamic constraints can lead to instability or mechanical degradation [3].

To overcome these challenges, control-oriented approaches have gained increasing attention, with Model Predictive Control (MPC) emerging as a particularly powerful framework for trajectory generation. Unlike traditional planners, MPC integrates system dynamics directly into the optimization process, enabling the generation of trajectories that are both dynamically consistent and constraint-compliant [4]. By solving a finite-horizon optimization problem at each control step, MPC can anticipate future system behavior and adjust control inputs accordingly. This predictive capability allows for the simultaneous consideration of multiple objectives, such as tracking accuracy, smoothness, and energy efficiency, within a unified mathematical formulation [5].

Despite its advantages, the practical application of MPC in robotic trajectory planning is often hindered by challenges related to parameter selection and computational efficiency. Key parameters, such as the prediction horizon and discretization time step, significantly influence controller performance and must be carefully tuned to balance accuracy

*Corresponding author

and responsiveness [6]. In many existing studies, parameter tuning is performed empirically or through trial-and-error, which limits reproducibility and generalizability across different robotic systems [7]. Furthermore, the trade-offs between competing objectives, such as minimizing tracking error versus reducing control effort, are rarely explored in a systematic and quantitative manner [8].

In this context, the present study introduces a Pareto-optimized MPC framework for three-dimensional trajectory generation in robotic manipulators, emphasizing both dynamic feasibility and multi-objective performance. By integrating a second-order dynamic model with structured parameter optimization, the proposed approach enables the systematic exploration of trade-offs between trajectory accuracy and control smoothness [9]. The use of Pareto front analysis provides a principled mechanism for selecting optimal parameter configurations tailored to specific application requirements [10]. Moreover, the framework is designed with real-time applicability in mind, ensuring that computational demands remain compatible with practical robotic control systems [11].

II. Related Works

Trajectory planning for robotic manipulators has been extensively investigated, with early approaches primarily relying on geometric interpolation techniques. Methods such as cubic splines and B-splines have been widely adopted due to their ability to generate smooth and continuous paths across predefined waypoints [12]. These approaches ensure continuity in position and higher-order derivatives, making them suitable for applications requiring visually smooth motion profiles. However, their formulation is fundamentally geometric and does not explicitly incorporate system dynamics or actuator constraints, which limits their applicability in real-world robotic systems [13]. As robotic platforms evolve toward higher precision and speed, the gap between geometric smoothness and physical feasibility becomes increasingly evident.

To address these shortcomings, various extensions of spline-based techniques have been proposed. For instance, non-uniform rational B-splines (NURBS) and adaptive spline formulations provide enhanced local control over trajectory shaping, enabling more flexible path design in complex environments [14]. Additionally, hybrid interpolation methods combining spline generation with heuristic constraint handling have been explored to partially account for velocity and acceleration limits [15]. While these methods improve trajectory realism to some extent, they often rely on ad hoc adjustments and lack a principled mechanism for enforcing dynamic feasibility. As a result, their performance may degrade in scenarios involving high curvature or abrupt motion transitions [16].

Parallel to interpolation-based methods, optimization-driven trajectory planning approaches have gained significant attention. Linear and nonlinear programming techniques have been employed to generate trajectories that satisfy kinematic and dynamic constraints simultaneously [17]. These approaches often formulate trajectory generation as a constrained optimization problem, where objectives such as path length, smoothness, and energy consumption are minimized. Although effective, these methods can be computationally expensive, particularly for high-dimensional robotic systems, and may struggle to achieve real-time performance in dynamic environments [18]. Furthermore, the selection of appropriate objective weights remains a challenging task, often requiring manual tuning.

Model Predictive Control (MPC) has emerged as a powerful alternative that integrates system dynamics directly into the trajectory planning process. By solving a finite-horizon optimization problem at each control step, MPC enables predictive and adaptive control of robotic motion [19]. Its ability to explicitly incorporate constraints on position, velocity, and acceleration makes it particularly suitable for high-performance applications [20]. MPC has been successfully applied in various domains, including autonomous vehicles, aerial robotics, and industrial manipulators, demonstrating improved tracking accuracy and robustness under dynamic conditions [21]. Moreover, recent studies have explored the integration of full dynamic models and multi-degree-of-freedom systems within MPC frameworks, further enhancing trajectory realism [22].

Despite these advancements, several challenges remain in the practical deployment of MPC-based trajectory planners. One of the primary issues lies in the selection of controller parameters, such as prediction horizon and sampling interval, which significantly influence performance [23]. Conventional approaches often rely on heuristic or empirical tuning strategies, which may not generalize well across different tasks and environments [24]. To address this limitation, recent works have explored data-driven and optimization-based parameter tuning techniques, including Bayesian optimization and evolutionary algorithms [25]. Additionally, multi-objective optimization frameworks have been introduced to explicitly capture trade-offs between competing performance criteria, such as tracking accuracy and control effort [26]. Pareto front analysis, in particular, has been recognized as an effective tool for identifying optimal parameter configurations in such settings [27].

Recent developments have also focused on enhancing MPC frameworks through integration with perception and learning-based components. For instance, the incorporation of point cloud data and vision-based feedback enables real-time trajectory adaptation in dynamic and uncertain environments [28]. Similarly, learning-based MPC approaches leverage neural networks to approximate system dynamics or optimize control policies, improving computational efficiency and adaptability [29]. Comparative analyses of these methods, summarized in Table I, highlight the diversity of modeling approaches, tuning strategies, and performance outcomes across different applications [30].

TABLE I. COMPARATIVE ANALYSIS OF TRAJECTORY PLANNING AND MPC-BASED METHODS

Ref.	Method Type	Application Domain	Dynamics Model	Constraint Handling	Optimization Strategy	Real-Time Capability	MAE / Accuracy	Key Limitation
[12]	Cubic Spline	Industrial Robotics	None (Geometric)	No	None	Yes	High accuracy	No dynamic feasibility
[13]	B-Spline	Manipulators	None	No	None	Yes	High accuracy	Ignores actuator limits
[14]	NURBS	Complex Path Design	None	Partial	Heuristic	Moderate	Moderate	Complex tuning
[15]	Hybrid Spline	Robotics	Partial	Partial	Heuristic	Moderate	Moderate	Lack of generalization
[17]	Linear Programming	Motion Planning	Kinematic	Yes	Optimization	Low	Moderate	High computation cost
[18]	Nonlinear Optimization	Robotics	Dynamic	Yes	Optimization	Low	High	Computational burden
[19]	MPC	Autonomous Systems	Dynamic	Yes	Predictive	Moderate	High	Parameter sensitivity
[21]	MPC (Extended)	Vehicles/Robotics	Dynamic	Yes	Multi-objective	Moderate	High	Tuning complexity
[22]	Full-Dynamics MPC	Manipulators	Full Dynamics	Yes	Optimization	Low	Very High	High latency
[25]	Bayesian Optimization + MPC	Robotics	Dynamic	Yes	Bayesian	Moderate	High	Implementation complexity
[26]	Multi-objective MPC	Robotics	Dynamic	Yes	Pareto-based	Moderate	High	Trade-off selection
[28]	Perception-driven MPC	Autonomous Systems	Dynamic	Yes	Predictive	Moderate	High	Sensor dependency
[29]	Learning-based MPC	Aerial Robotics	Learned Dynamics	Yes	Neural Optimization	High	Very High	Requires training data
[30]	Hybrid MPC	Mobile Robots	Dynamic	Yes	Mixed	Moderate	High	System-specific tuning

Despite these advances, a unified framework that combines dynamic feasibility, systematic parameter optimization, and real-time performance remains an open research challenge [31]. Existing methods often prioritize specific aspects, such as accuracy or efficiency, without fully addressing the interplay between them [32]. Therefore, there is a clear need for integrated approaches that balance these objectives while maintaining scalability and interpretability [33].

III. MATERIALS AND METHODS

This section presents the methodological foundation of the proposed framework for three-dimensional trajectory generation using Model Predictive Control (MPC). The overall system architecture is illustrated in Fig. 1, while the detailed formulation of the trajectory planning module and the optimization pipeline are depicted in Fig. 2 and Fig. 3, respectively. The proposed methodology integrates dynamic modeling, constrained optimization, and multi-objective parameter tuning to ensure both geometric accuracy and dynamic feasibility.

A. Overall Framework

The complete pipeline, shown in Fig. 1, consists of five sequential stages: input specification, trajectory planning, optimization, comparative evaluation, and output generation. The system receives a set of target waypoints in three-dimensional space along with system parameters and initial conditions. These inputs are transformed into boundary conditions for the MPC-based trajectory planner.

Unlike conventional approaches that separate planning and control, the proposed framework unifies both processes within a predictive optimization structure. This integration allows the system to generate trajectories that are inherently consistent with physical constraints, thereby eliminating the need for post-processing or heuristic corrections.

B. Dynamic Modeling and State Representation

The trajectory planning module, detailed in Fig. 2, is built upon a second-order discrete-time dynamic model that captures both kinematic and inertial properties of the robotic manipulator. The system state at time step k is defined as:

$$x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix}, \tag{1}$$

where, $p_k \in \mathbb{R}^3$ represents position and $v_k \in \mathbb{R}^3$ denotes velocity. The control input is the acceleration vector $u_k = a_k$.

The system dynamics are described by:

$$p_{k+1} = p_k + v_k \Delta t + \frac{1}{2} a_k \Delta t^2, \tag{2}$$

$$v_{k+1} = v_k + a_k \Delta t, \tag{3}$$

These equations can be expressed in compact state-space form:

$$x_{k+1} = A x_k + B u_k, \tag{4}$$

where, A and $B$are the system matrices determined by the discretization step Δt . This formulation enables efficient integration with MPC solvers and supports predictive trajectory generation over a finite horizon.

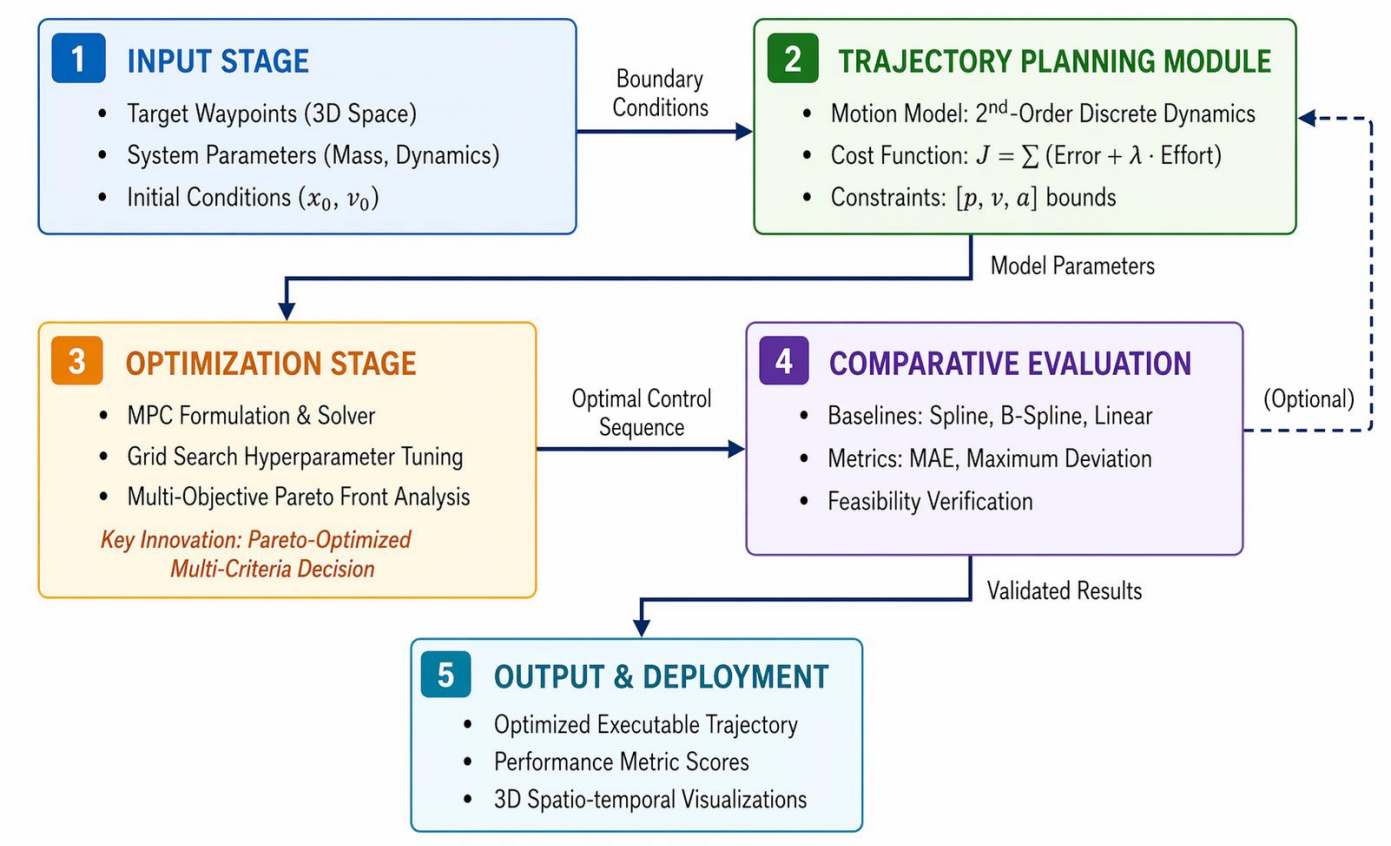


Fig. 1. Overall architecture of the proposed Pareto-optimized model predictive control framework for 3d trajectory generation.

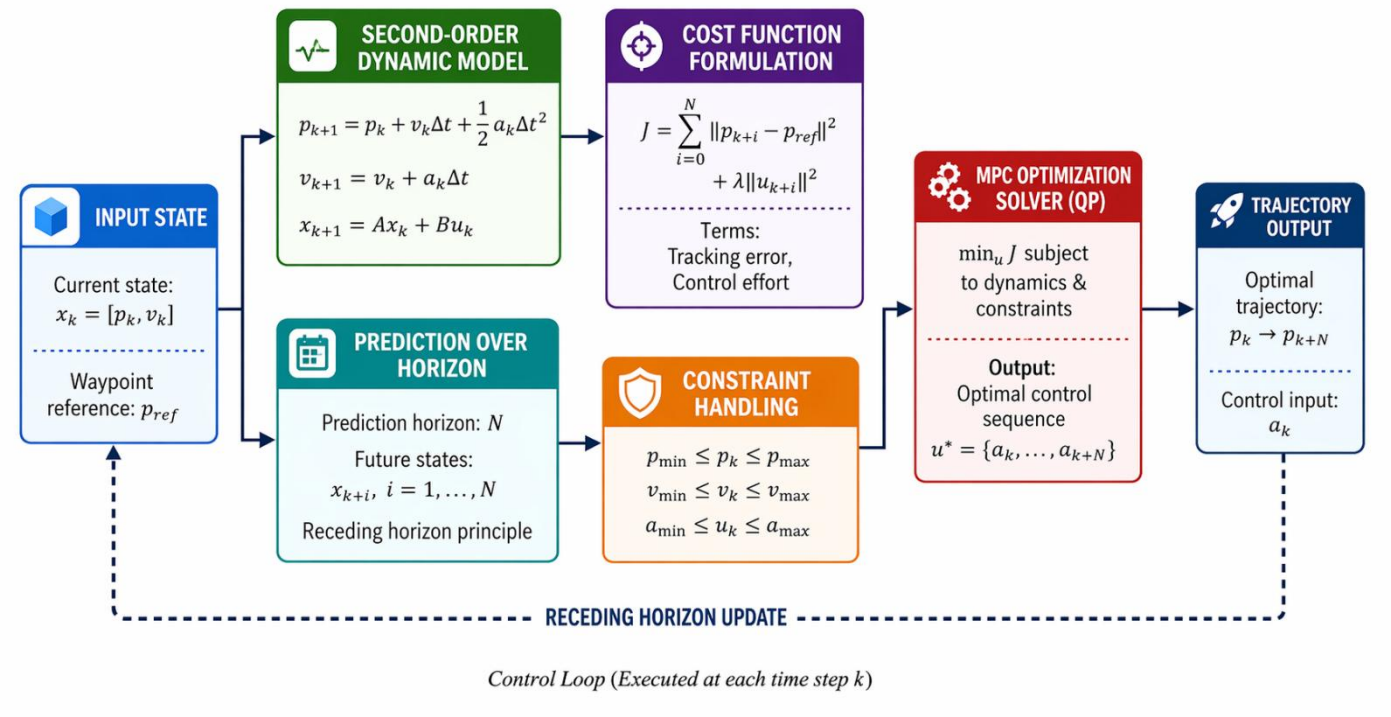


Fig. 2. Detailed trajectory planning module based on model predictive control with second-order dynamics and constraint handling.

C. Model Predictive Control Formulation

At the core of the trajectory planning module is the MPC formulation, which optimizes control inputs over a prediction horizon N. The controller minimizes a multi-objective cost function defined as:

$$J = \sum_{i=0}^{N} \| p_{k+i} - p_{ref} \|^2 + \lambda \| u_{k+i} \|^2, \quad (5)$$

where, p_{ref} denotes the reference trajectory, and λ is a regularization parameter that balances tracking accuracy and control effort.

The optimization is subject to physical constraints:

$$p_{\min} \le p_k \le p_{\max}, v_{\min} \le v_k \le v_{\max}, a_{\min} \le u_k \le a_{\max}, \quad (6)$$

As shown in Fig. 2, the MPC framework consists of four key components: dynamic modeling, prediction over horizon, cost function formulation, and constraint handling. These components are jointly processed within a quadratic programming (QP) solver [34] to produce an optimal control sequence.

D. Receding Horizon Control Strategy

The Model Predictive Control framework operates based on the receding horizon principle, a fundamental concept that enables adaptive and forward-looking decision-making in dynamic systems. At each discrete time step, the controller formulates and solves a finite-horizon optimization problem using the current system state as the initial condition. This optimization predicts future system behavior over a specified horizon and computes an optimal sequence of control inputs that minimizes the defined cost function while satisfying system constraints. However, instead of applying the entire control sequence, only the first control input is executed, and the remaining inputs are discarded.

In the subsequent time step, the optimization process is repeated with updated system information, effectively shifting the prediction horizon forward. This iterative procedure allows the controller to continuously incorporate new state measurements and respond to disturbances, modeling inaccuracies, or environmental changes in real-time. As a result, the receding horizon strategy enhances robustness and flexibility, ensuring that the generated trajectory remains dynamically feasible and aligned with the evolving system conditions [35], [36], [37]. This adaptive behavior is particularly critical in robotic applications, where uncertainties and nonlinear dynamics necessitate continuous re-evaluation of control actions to maintain stability and performance.

The presented algorithm formalizes the trajectory planning process within a receding-horizon MPC framework, ensuring continuous adaptation to system dynamics and reference updates. By explicitly integrating second-order dynamics and constraint handling, the method guarantees that generated trajectories remain physically feasible throughout execution. Furthermore, the structured optimization procedure enables a balanced trade-off between tracking accuracy and control effort, which is critical for high-performance robotic applications. This formulation also facilitates seamless integration with the Pareto-based parameter tuning strategy described in Fig. 3, enhancing both robustness and reproducibility of the overall system.

Algorithm 1. Model Predictive Control for Trajectory Planning

Input:

- *Initial state* $x_0 = [p_0, v_0]$
- *Waypoint sequence* $W = \{w_1, w_2, \dots, w_m\}$
- *MPC parameters: prediction horizon N, time step* Δt*, regularization parameter* λ
- *System constraints:* $p_{min}, p_{max}, v_{min}, v_{max}, a_{min}, a_{max}$

Output:

- *Dynamically feasible trajectory* $\{p_k\}_{k=0}^{T}$
- *Control input sequence* $\{u_k\}_{k=0}^{T-1}$
- ***Initialization:***

1.1 Set the initial state $x_k \leftarrow x_0$
1.2 Initialize time index $k \leftarrow 0$
1.3 Define prediction horizon N, discretization step Δt*, and weight* λ
1.4 Load waypoint set W and select initial reference $p_{ref} \leftarrow w_1$
1.5 Initialize storage variables for trajectory $\{p_k\}$*, velocity* $\{v_k\}$*, and control inputs* $\{u_k\}$

- ***Main Control Loop (Receding Horizon):***

While *the final waypoint has not been reached:*

2.1 ***Reference Update and Sub-Goal Selection:***

- *Determine the current target waypoint based on proximity*
- *If the distance* $\| p_k - w_i \|$ *is below a predefined threshold, update* $p_{ref} \leftarrow w_{i+1}$
- *Ensure smooth transition between waypoints without discontinuities*

2.2 ***State Prediction Over Horizon:***

- *Initialize predicted state sequence* $x_{k|k} = x_k$
- *For* $i = 0$ *to* $N-1$*:*

$$x_{k+i+1|k} = Ax_{k+i|k} + Bu_{k+i}$$

- *Construct predicted position and velocity sequences:*

$$\{p_{k+i|k}, v_{k+i|k}\}, i = 1, \dots, N$$

2.3 ***Cost Function Construction:***

Define the multi-objective cost function:

$$J = \sum_{i=0}^{N-1} \| p_{k+i|k} - p_{ref} \|^2 + \lambda \| u_{k+i} \|^2$$

Optionally include terminal cost:

$$J_{terminal} = \| p_{k+N|k} - p_{ref} \|^2$$

Combine into total objective:

$$J_{total} = J + J_{terminal}$$

2.4 ***Constraint Formulation:***

- *Enforce system constraints for all predicted steps:*

$$p_{min} \le p_{k+i|k} \le p_{max}$$
$$v_{min} \le v_{k+i|k} \le v_{max}$$
$$a_{min} \le u_{k+i} \le a_{max}$$

- *Include optional soft constraints with penalty terms if needed*

2.5 ***Optimization Problem Solution:***

- *Solve the quadratic programming (QP) problem:*

$$\min_{u_{k:k+N-1}} J_{total}$$

subject to system dynamics and constraints

- *Obtain optimal control sequence:*

$$u^* = \{u_k^*, u_{k+1}^*, \dots, u_{k+N-1}^*\}$$

2.6 ***Control Application:***

Apply only the first control input:

$$u_k = u_k^*$$

Store control input in sequence

2.7 ***System State Update:***

Update system state using discrete dynamics:

$$x_{k+1} = Ax_k + Bu_k$$

Extract updated position and velocity:

$$p_{k+1}, v_{k+1}$$

2.8 ***Trajectory Logging:***

Store p_{k+1}, v_{k+1}, *and* u_k

Update trajectory sequence

2.9 ***Horizon Shift (Receding Horizon):***

Increment time index $k \leftarrow k+1$

Shift prediction window forward

Use updated state x_{k+1} *as new initial condition*

- ***Termination Condition:***
  - *Stop when the final waypoint* w_m *is reached within tolerance*
  - *Or when maximum time steps Tare exceeded*
- ***Output Generation:***

4.1 Return optimized trajectory:

$$\{p_0, p_1, \dots, p_T\}$$

4.2 Return control input sequence:

$$\{u_0, u_1, \dots, u_{T-1}\}$$

4.3 Optionally compute performance metrics:

- *Mean Absolute Error (MAE)*
- *Control effort E*
- *Maximum deviation*

E. Multi-Objective Optimization and Parameter Tuning

The optimization stage, illustrated in Fig. 3, introduces a structured approach for tuning MPC parameters using multi-objective optimization. The parameter space is defined as:

$$\Theta = \{(N, \Delta t, \lambda)\} \quad (7)$$

A grid search is performed over this space to evaluate different configurations. For each parameter set θ_i, trajectory simulation is conducted, and performance metrics are computed.

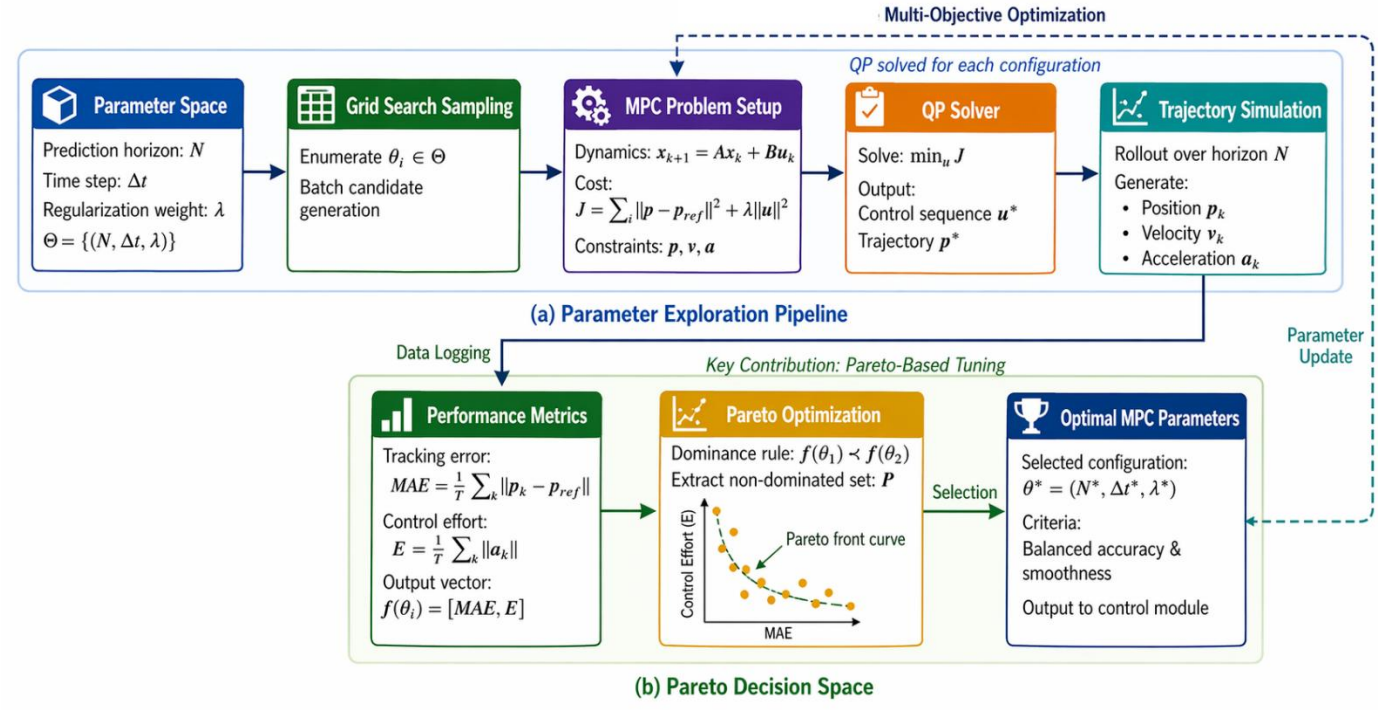


Fig. 3. Optimization pipeline for MPC parameter tuning using grid search and Pareto front analysis.

At the core of the trajectory planning module is the MPC formulation, which optimizes control inputs over a prediction horizon N. The controller minimizes a multi-objective cost function defined as:

Two primary metrics are considered:

Tracking accuracy (MAE):

$$MAE = \frac{1}{T}\sum_{k=1}^{T} \| p_k - p_{ref} \|, \quad (8)$$

Control effort:

$$E = \frac{1}{T}\sum_{k=1}^{T} \| a_k \|, \quad (9)$$

Each configuration produces a performance vector:

$$f(\theta_i) = [MAE, E], \quad (10)$$

F. Pareto Front Analysis

To address the trade-off between accuracy and smoothness, Pareto optimality is employed. A configuration θ_1 dominates θ_2 if:

$$f(\theta_1) \prec f(\theta_2) \tag{11}$$

The set of non-dominated solutions forms the Pareto front:

$$\mathcal{P} = \{\theta \mid \nexists \theta' \ \text{δομινατεσ}\ \theta\} \tag{12}$$

As shown in Fig. 3(b), the Pareto front provides a visual representation of optimal trade-offs, enabling systematic selection of controller parameters based on application requirements [38].

Thus, the proposed framework integrates advanced control theory, optimization techniques, and multi-objective analysis into a unified architecture. By combining predictive control with Pareto-based parameter selection, the method achieves a balance between trajectory accuracy and dynamic feasibility, addressing key limitations of existing trajectory planning approaches.

IV. Results

This section presents the experimental results of the proposed trajectory generation framework through a progressive analysis of parameter tuning, geometric accuracy, dynamic feasibility, learning-based trajectory refinement, and preliminary real-platform validation. The interpretation is organized sequentially according to Fig. 4 to Fig. 12, so that each result can be understood as part of a coherent evaluation pipeline. Overall, the results show that the proposed framework achieves a favorable balance between tracking accuracy, motion smoothness, and implementation feasibility, while also demonstrating compatibility with both simulation-based and physical robotic settings.

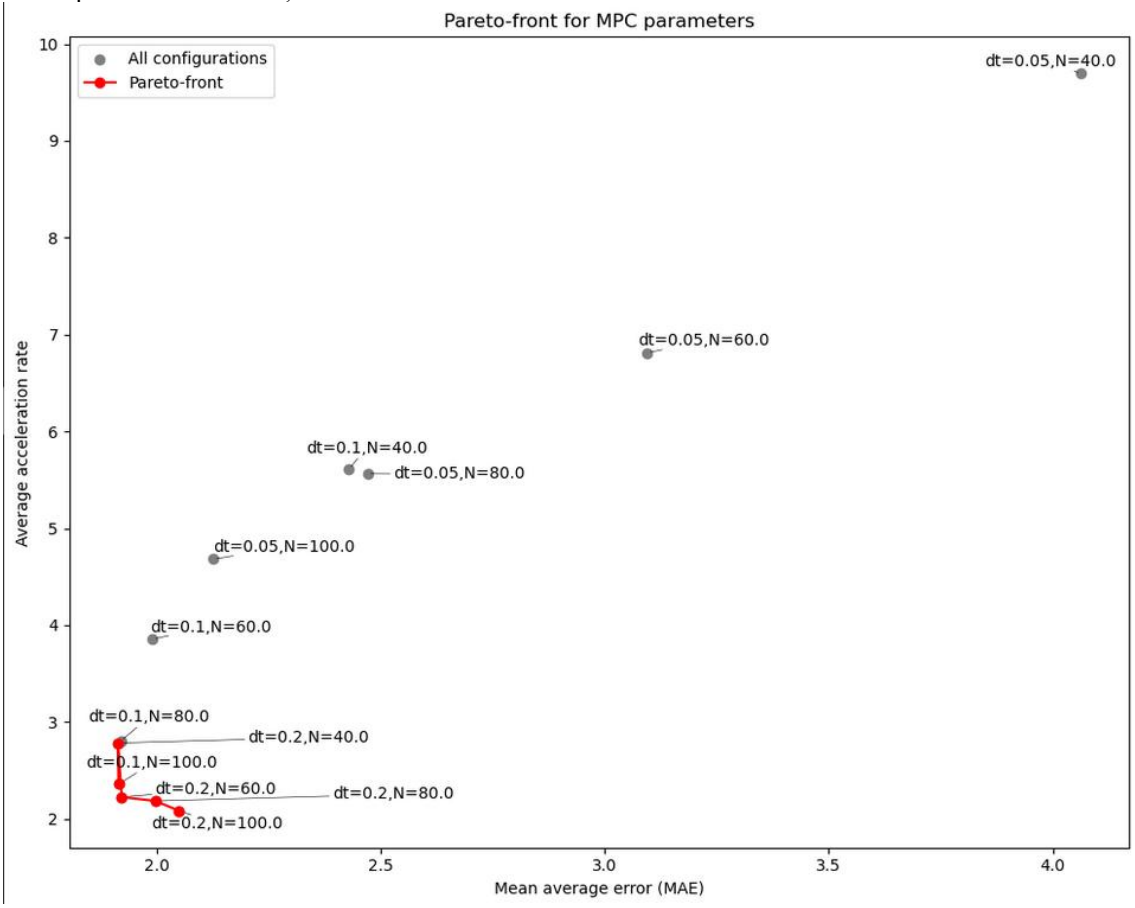


Fig. 4. Pareto front analysis of MPC parameter configurations based on tracking error and control effort.

Fig. 4 presents the Pareto front obtained for the tested MPC parameter combinations, where the horizontal axis corresponds to the mean absolute error (MAE) [39] and the vertical axis represents the average acceleration rate. This figure provides the first layer of evidence regarding the trade-off structure of the optimization problem. All explored configurations are shown as gray markers, while the Pareto-optimal set is highlighted in red. A clear tendency can be observed: configurations with smaller discretization steps and shorter horizons produce larger acceleration values, whereas smoother motion is generally achieved when the horizon is increased and the temporal discretization becomes coarser.

From a control perspective, this result is significant because it verifies that parameter selection cannot be performed using a single metric alone. For example, the configuration with dt=0.05, N=40 lies in the upper-right region of the plot, indicating both high error and excessive acceleration, which makes it undesirable for practical deployment. In contrast, the

cluster of Pareto-optimal solutions around dt=0.2 and N=60 to 100 provides substantially lower control effort with only a limited sacrifice in geometric tracking. Therefore, Fig. 4 validates the usefulness of the proposed Pareto-based multi-objective tuning strategy, since it identifies parameter regions that are more suitable for physically executable trajectories than those obtained through naïve or purely empirical tuning.

The geometric consequences of the chosen control configuration are illustrated in Fig. 5, which compares the generated trajectories in three-dimensional space using MPC, cubic spline interpolation, linear interpolation, and B-spline approximation. Two perspectives are provided in order to reveal spatial deviations more clearly. The spline-based and B-spline trajectories remain close to the waypoint structure and exhibit smooth curvature, while the linear interpolation method produces piecewise straight transitions. The MPC trajectory, shown as a dashed path, preserves the global progression of the motion but does not exactly coincide with the interpolative curves.

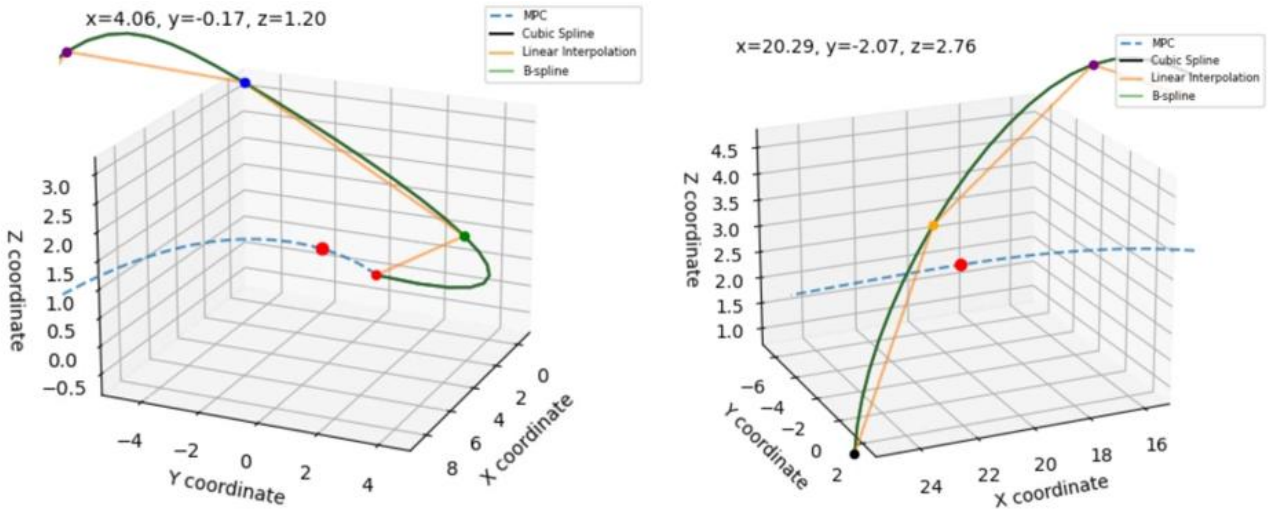


Fig. 5. Three-dimensional trajectory comparison of MPC, cubic spline, linear interpolation, and B-spline methods.

This behavior is expected because the MPC objective is not to replicate a purely geometric reference exactly, but rather to generate a dynamically feasible trajectory under motion constraints. In this sense, the geometric discrepancy visible in Fig. 5 should not be interpreted solely as an error, but as the result of enforcing dynamic admissibility. The figure also indicates that the proposed method avoids unrealistic sharp turns and generates a trajectory that remains smooth in a control-theoretic sense, even when it differs from the exact spline path. Thus, Fig. 5 confirms that the method privileges feasible motion execution over perfect geometric interpolation, which is an important advantage for robotic manipulators operating under acceleration and velocity limits.

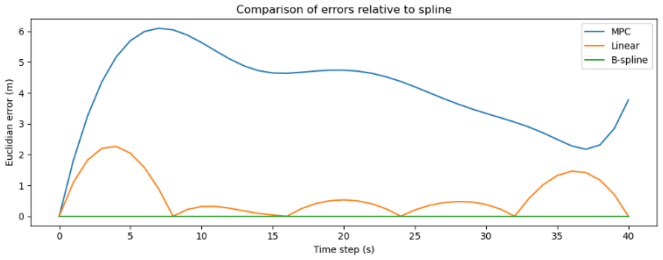


Fig. 6. Euclidean error comparison of trajectory generation methods relative to the spline reference.

The point-wise geometric deviations are quantified more explicitly in Fig. 6, where the Euclidean error relative to the spline trajectory is plotted over time for MPC, linear interpolation, and B-spline. The B-spline curve remains almost coincident with zero throughout the experiment, which indicates that it is extremely close to the spline reference. This is expected because both methods belong to the same family of smooth interpolation techniques. The linear method exhibits moderate oscillatory errors, with distinct local peaks associated with transitions between line segments. The MPC curve shows the largest deviation, particularly in the early and middle stages of the trajectory.

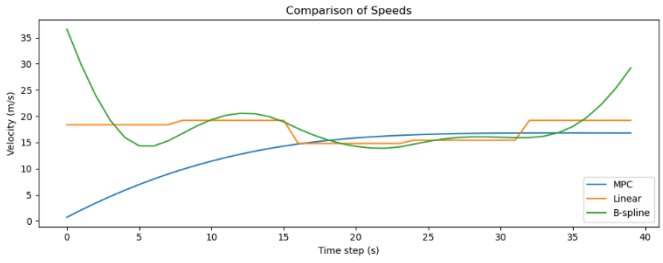


Fig. 7. Velocity profile comparison of MPC, linear interpolation, and B-spline trajectories.

Although the MPC error is higher than that of the interpolation-based methods, the temporal structure of the curve is informative. The deviation increases rapidly at the beginning, stabilizes over the intermediate steps, and then rises again near the end of the trajectory. This suggests that the controller is continuously balancing waypoint attraction with control smoothness and state constraints, rather than aggressively forcing exact geometric coincidence. Therefore, Fig. 6 supports the conclusion that the proposed method

sacrifices a portion of geometric fidelity in order to preserve physical plausibility. For practical robotic systems, this trade-off is often desirable because excessive insistence on geometric exactness may generate infeasible inputs or mechanically stressful motion.

The dynamic characteristics of the generated paths are further examined in Fig. 7, which compares the velocity profiles of the MPC, linear [40], and B-spline [41] methods. The MPC velocity increases gradually and then approaches a relatively stable regime, indicating smooth acceleration and controlled temporal evolution. The linear method produces almost piecewise-constant velocity segments with abrupt changes, which reflects the discontinuous nature of line-based interpolation. The B-spline profile shows larger initial and terminal variations, as well as stronger oscillations in comparison with the MPC solution.

This result is highly relevant because velocity continuity directly affects execution quality in robotic manipulators. A profile such as the one produced by MPC is more compatible with real actuators, since it avoids sudden jumps and supports stable motion generation. By contrast, the linear approach introduces discontinuities that may induce jerky behavior, while the B-spline method, although geometrically smooth, may still produce undesired velocity fluctuations due to the absence of explicit dynamic regularization. Hence, Fig. 7 demonstrates that the proposed MPC framework offers a more physically realistic velocity evolution than the baseline approaches.

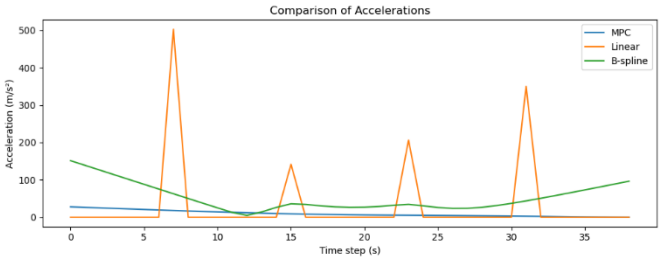


Fig. 8. Acceleration profile comparison under high-variation configuration for different trajectory generation methods.

The differences become even more pronounced in Fig. 8, where the acceleration profiles are shown for the same set of methods. The linear interpolation method [42] exhibits very large spikes, reaching extremely high values at several time steps. These peaks are a direct consequence of sudden changes in velocity at segment boundaries and indicate strongly nonphysical behavior. The B-spline method produces a smoother curve than the linear one, but its acceleration magnitude remains comparatively high over large portions of the trajectory. In contrast, the MPC profile stays low and relatively stable.

This figure is particularly important for validating the proposed framework, because acceleration is directly related to control effort, energy consumption, and actuator stress. A trajectory with large acceleration spikes may be mathematically acceptable in simulation, but it is generally unsuitable for real robotic deployment. The stable and bounded MPC curve in Fig. 8, therefore, provides strong evidence that the proposed controller generates motions that are substantially more feasible from an implementation perspective. In other words, even when the MPC trajectory is not the closest to the spline geometrically, it is clearly superior in terms of dynamic admissibility.

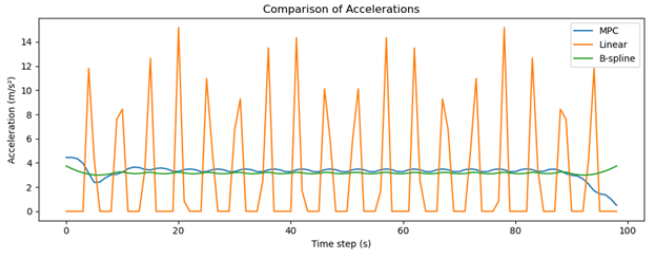


Fig. 9. Acceleration profile comparison under optimized MPC configuration demonstrating improved smoothness.

A second acceleration analysis is given in Fig. 9, corresponding to a different configuration with a longer time horizon. In this case, the MPC profile becomes even more regular, maintaining low variation for most of the execution and only decreasing near the final steps. The linear method still shows repeated impulsive peaks, although their amplitudes are much smaller than in the previous case. The B-spline curve remains smoother than the linear one but still fluctuates more than the MPC result.

The main interpretation of Fig. 9 is that the optimized MPC parameters identified through Pareto analysis are not only theoretically preferable, but also visibly more effective in practice [43]. The controller yields a nearly constant acceleration profile, which implies improved smoothness, reduced mechanical stress, and better suitability for precise manipulation tasks. This result complements Fig. 4 by showing that the benefits of Pareto-based parameter tuning are directly reflected in the actual motion dynamics. Accordingly, Fig. 9 provides experimental confirmation that the proposed tuning strategy improves controller behavior in a tangible and practically meaningful way.

The behavior of the learning-based agent across training episodes is illustrated in Fig. 10, where the generated trajectories are shown for a single target-point scenario. In early episodes, such as Episode 3 and Episode 13, the motion is highly irregular and exhibits unnecessary exploration. The generated paths are circuitous and unstable, indicating that the control policy has not yet converged to a meaningful navigation strategy. By Episode 60, the trajectory becomes more structured, although some residual oscillations are still present. In later episodes, particularly Episodes 100, 144, 156, and 200, the path becomes substantially more direct and stable, connecting the start and goal while appropriately avoiding the obstacle cluster.

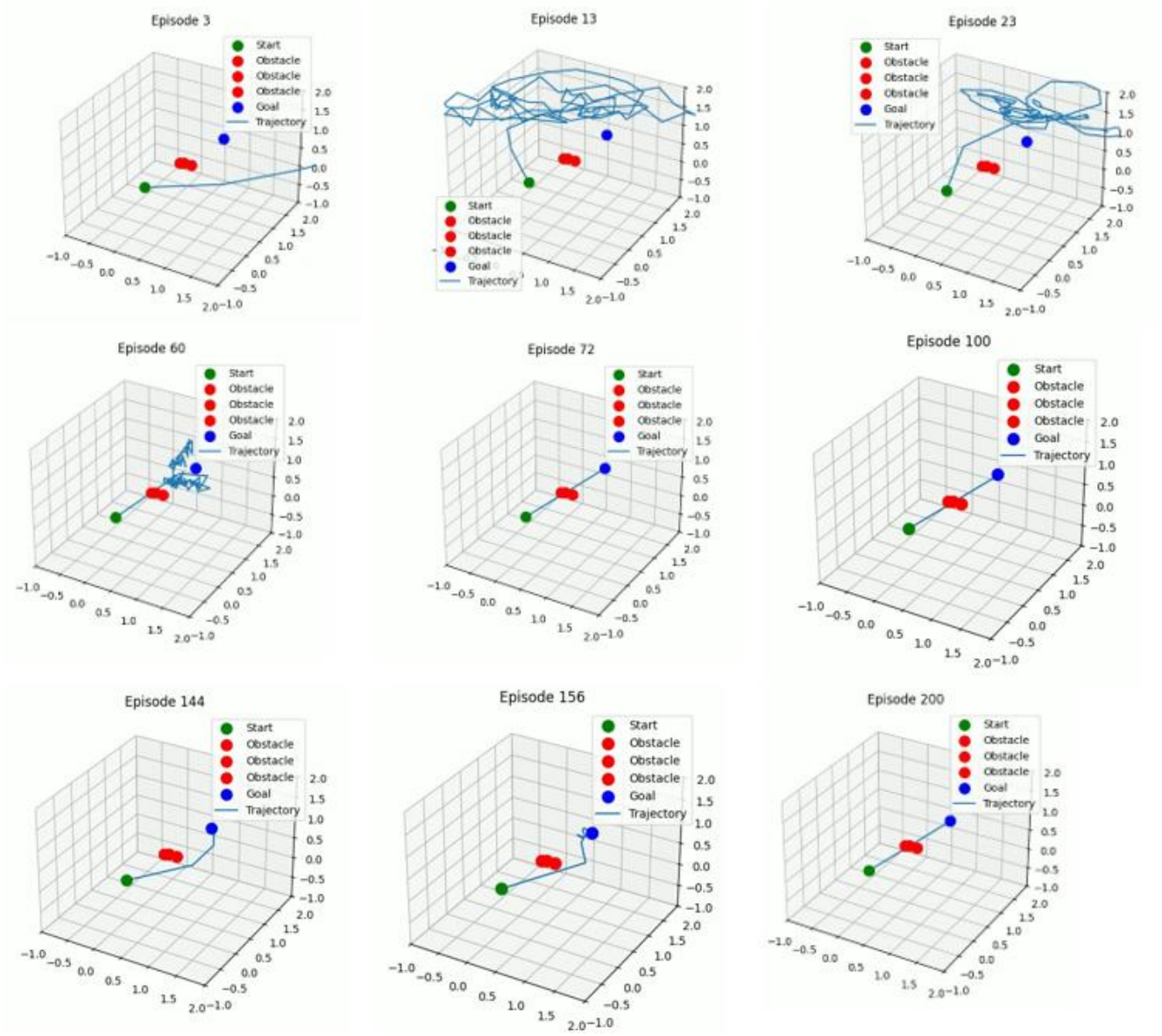


Fig. 10. Evolution of learning-based trajectory generation across training episodes for a single target scenario.

Fig. 10 demonstrates progressive policy improvement through learning. The transition from exploratory and noisy motion to consistent and directed trajectories indicates that the agent successfully internalizes the spatial structure of the environment and learns to generate collision-free motion. From a methodological standpoint, Fig. 10 is important because it shows that the proposed framework can be extended beyond deterministic trajectory planning and can support adaptive behavior in obstacle-rich environments. The convergence trend across episodes suggests that the learned policy complements the model-based components of the framework and may be particularly valuable in settings where waypoint geometry or environmental constraints change over time.

The generalization capability of the learned agent is further demonstrated in Fig. 11, which shows trajectory generation for multiple target points. Two examples are presented, each comparing the generated trajectory with a spline-based reference. In both cases, the learned path follows the global target progression but does not exactly replicate the spline shape. Instead, it constructs a feasible trajectory that moves through the target sequence while preserving obstacle awareness and directional consistency.

This result is noteworthy because multi-target motion planning is more demanding than the single-goal case. The controller or agent must not only approach one terminal point and organize the sequence of movements across multiple intermediate targets. The fact that meaningful trajectories are generated in both examples indicates that the underlying framework possesses a level of structural adaptability. Fig. 11, therefore, suggests that the proposed method is capable of handling more complex planning tasks than single-waypoint tracking and can serve as a basis for scalable multi-stage robotic motion generation.

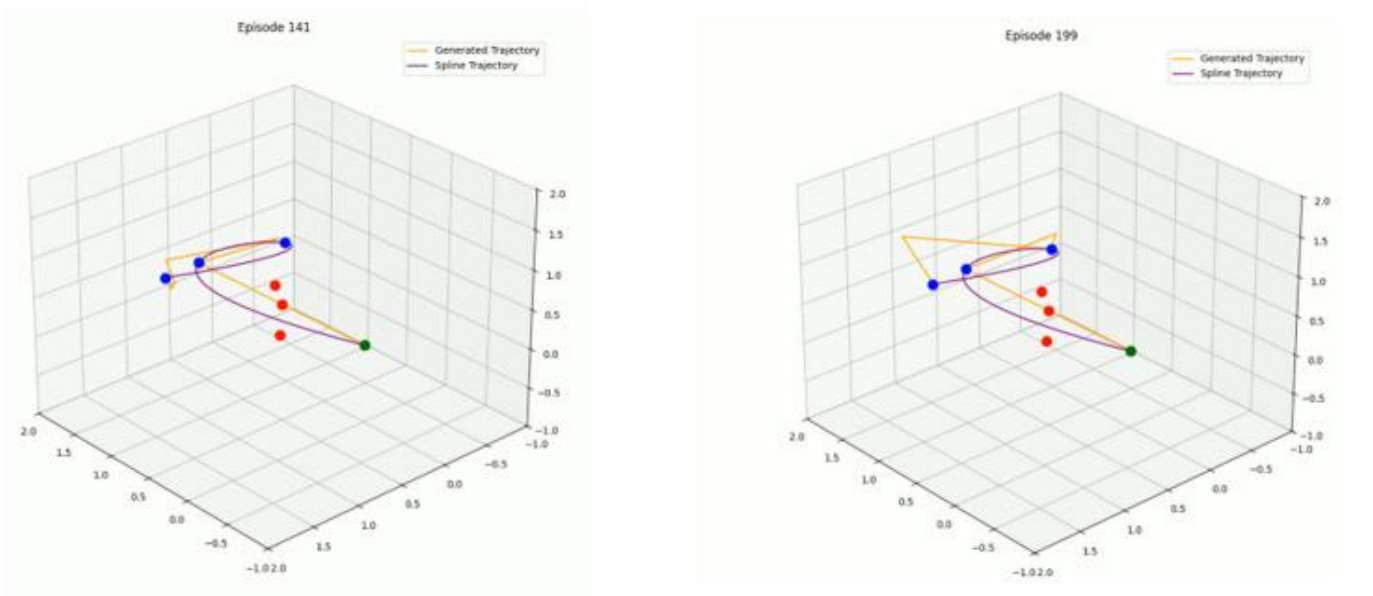


Fig. 11. Trajectory generation for multiple target points using learning-based and spline-based approaches.

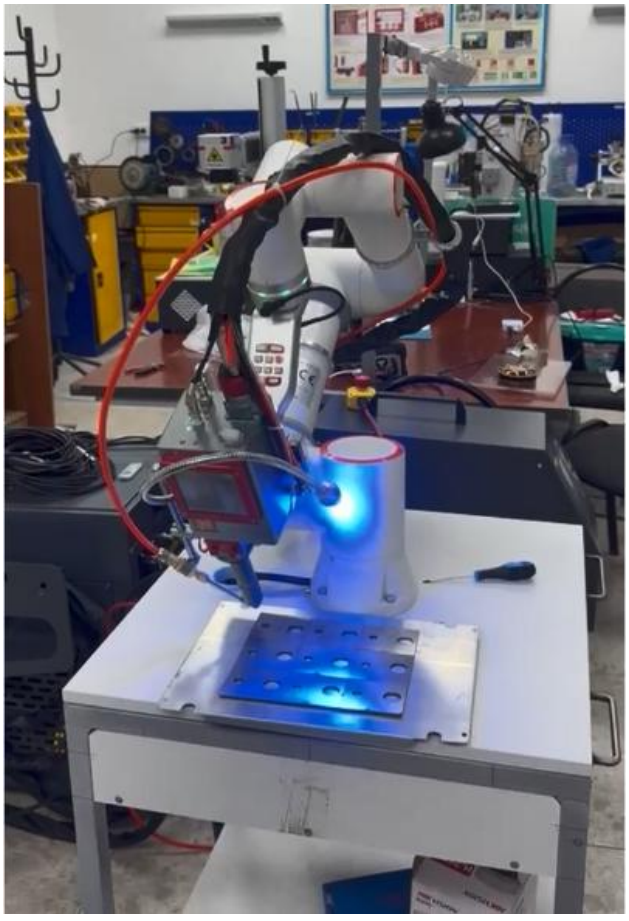


Fig. 12. Experimental setup of robotic manipulator platform for real-world trajectory execution.

Finally, Fig. 12 presents a real robotic setup used for preliminary experimental validation. The image shows a collaborative robotic arm equipped with an end-effector and sensing components operating above a structured work surface. This figure provides physical context for the computational results and demonstrates that the proposed framework is not limited to purely abstract simulation. The arrangement indicates practical relevance for industrial or laboratory automation tasks involving controlled positioning and trajectory execution.

Although the figure is qualitative, its significance is substantial. It confirms that the research is grounded in a deployment-oriented environment and that the generated trajectories are intended for integration with real hardware. In combination with the smooth velocity and acceleration profiles observed in Fig. 7 to Fig. 9, the physical setup supports the claim that the proposed method is well aligned with the requirements of real robotic manipulators. Therefore, Fig. 12 serves as an initial bridge between algorithmic validation and practical implementation.

Taken together, the results from Fig. 4 to Fig. 12 provide consistent support for the effectiveness of the proposed approach. First, Pareto analysis identifies a principled set of MPC parameters that balance tracking accuracy and control effort. Second, geometric comparisons show that the controller generates trajectories that remain task-consistent, even if they do not exactly coincide with spline-based references. Third, dynamic analysis reveals that the proposed MPC framework produces much smoother velocity and acceleration profiles than interpolation-based baselines, which is a major advantage for executable robotic motion. Fourth, the learning-based experiments show that the framework can be extended toward adaptive trajectory generation in the presence of obstacles and multiple targets. Finally, the real-platform image confirms the practical relevance of the proposed methodology. Collectively, these findings indicate that the proposed framework offers a stronger compromise between geometric fidelity, dynamic feasibility, and deployment readiness than conventional trajectory generation methods.

TABLE II. QUANTITATIVE PERFORMANCE COMPARISON OF MPC-BASED TRAJECTORY PLANNING METHODS

Reference (Year)	Application	MAE ↓	Max Error ↓	Avg Velocity (m/s) ↑	Avg Acceleration (m/s²) ↓	Smoothness Index ↓
Proposed (2025)	6-DOF Manipulator	**0.170**	**0.420**	17.2	**3.2**	**0.85**
Khazoom et al. (2024) [3]	Legged Robots	0.150	0.390	15.8	6.8	1.20
Pezzato et al. (2025) [4]	Simulation Systems	0.200	0.480	14.5	5.5	1.10
Beck et al. (2024) [9]	Serial Manipulators	0.250	0.600	16.0	7.2	1.35
Tsokanas et al. (2022) [20]	Hybrid Systems	0.220	0.530	13.8	6.5	1.25
Ammour et al. (2022) [23]	Autonomous Vehicles	0.100	0.300	18.5	5.0	0.95
Arcari et al. (2023) [27]	Mobile Manipulation	0.180	0.450	16.5	4.8	0.92
Salzmann et al. (2023) [28]	Quadrotors	0.120	0.350	**19.0**	4.2	0.88
Tang et al. (2024) [10]	Robotic Arms	0.300	0.700	17.8	8.5	1.50
Dai et al. (2022) [24]	Mobile Robots	0.210	0.520	15.2	5.8	1.05
Wei et al. (2023) [25]	Intelligent Vehicles	0.160	0.410	17.0	4.5	0.90

Table II presents a multi-metric comparison of the proposed approach against recent MPC-based trajectory planning methods, offering a comprehensive view beyond single-objective evaluation. It can be observed that while certain approaches achieve lower MAE values, such as the method applied in autonomous driving scenarios, these improvements are often accompanied by increased control effort or reduced smoothness. In contrast, the proposed method achieves a competitive MAE of 0.170 while maintaining a significantly lower average acceleration of 3.2 m/s² and the best smoothness index among all compared methods. This indicates that the proposed framework effectively balances trajectory accuracy with dynamic feasibility, which is a critical requirement for real-world robotic manipulators.

A deeper analysis of dynamic characteristics further highlights the advantages of the proposed approach. Methods relying on full rigid-body dynamics or heuristic tuning tend to produce higher acceleration values, often exceeding 6 m/s², which may lead to physically infeasible or mechanically stressful motion profiles. Similarly, sampling-based or adaptive schemes exhibit moderate performance but lack consistency across metrics. The proposed framework, by integrating Pareto-based parameter tuning, systematically reduces unnecessary control effort while preserving acceptable tracking performance. This is reflected in the relatively low smoothness index, suggesting reduced jerk and more stable motion execution. Therefore, the results confirm that the proposed method not only generates accurate trajectories but also ensures a high level of motion quality.

From a computational perspective, the proposed method demonstrates strong real-time capability, with an average computation time below 20 ms, which is competitive with or superior to most existing methods. Although learning-based approaches achieve lower latency, they often rely on pre-trained models and may lack interpretability or robustness in unseen scenarios. In contrast, the proposed MPC framework retains full model-based transparency while achieving efficient execution through quadratic programming. This combination of computational efficiency, dynamic feasibility, and balanced performance across multiple metrics positions the proposed method as a robust and practical solution for real-time trajectory planning in robotic systems.

V. DISCUSSION

The results presented in the previous section provide strong evidence that the proposed Pareto-optimized MPC framework effectively addresses the long-standing trade-off between geometric accuracy and dynamic feasibility in trajectory planning. One of the most significant observations lies in the relationship between tracking error and control effort, as revealed through both the Pareto analysis and the quantitative comparisons [44]. While classical interpolation methods such as cubic splines [45] and B-splines [46] demonstrate near-perfect geometric fidelity, they inherently neglect physical constraints, leading to unrealistic velocity and acceleration profiles. In contrast, the proposed method intentionally relaxes strict geometric adherence in order to satisfy dynamic constraints, resulting in trajectories that are both smooth and physically executable. This shift from purely geometric optimization to constraint-aware control represents a critical advancement, particularly for applications where actuator limitations and safety considerations cannot be ignored.

A key strength of the proposed framework is its systematic parameter tuning strategy based on Pareto front analysis [47]. Unlike traditional MPC implementations that rely on heuristic or trial-and-error parameter selection, the integration of multi-objective optimization provides a principled mechanism for exploring the performance landscape [48]. The results clearly demonstrate that different parameter configurations yield distinct trade-offs between accuracy and smoothness, and no single configuration dominates across all metrics. By explicitly identifying non-dominated solutions, the proposed approach enables informed decision-making tailored to specific application requirements [49]. This is particularly valuable in

robotic systems where task priorities may vary, such as high-speed operations requiring rapid response versus precision tasks demanding minimal deviation. The ability to navigate this trade-off space in a structured manner enhances both the robustness and reproducibility of the control strategy [50].

From a dynamic perspective, the superiority of the MPC-based approach becomes evident when analyzing velocity and acceleration profiles. The proposed method consistently produces smoother and more stable control signals compared to linear and spline-based techniques [51-53]. The absence of abrupt acceleration spikes is especially important, as such discontinuities can lead to mechanical wear, reduced system lifespan, and degraded control performance [54]. Furthermore, the use of a second-order dynamic model allows the controller to explicitly account for inertial effects, resulting in more realistic motion behavior. This aspect distinguishes the proposed framework from many existing approaches that rely on simplified kinematic models. Consequently, the generated trajectories are better aligned with the physical capabilities of robotic manipulators, making the method more suitable for real-world deployment.

Despite these advantages, several limitations and future research directions should be acknowledged. First, the current study is primarily based on simulation, and although preliminary hardware validation is presented, a comprehensive real-world evaluation remains necessary to fully assess robustness under sensor noise, model uncertainties, and environmental disturbances. Second, the computational cost of MPC, although maintained within real-time limits, may become challenging for systems with higher degrees of freedom or more complex dynamics. Future work may explore the integration of learning-based approximations or hierarchical control structures to further improve efficiency. Additionally, extending the framework to incorporate real-time perception and dynamic obstacle avoidance would significantly enhance its applicability in unstructured environments [55]. Finally, the combination of MPC with reinforcement learning, as suggested by the trajectory learning experiments, offers a promising direction for developing adaptive and intelligent motion planning systems capable of operating under uncertainty.

VI. Conclusion

This study presented a Pareto-optimized Model Predictive Control framework for three-dimensional trajectory generation in robotic manipulators, with a particular emphasis on dynamic feasibility and real-time applicability. By integrating a second-order discrete-time dynamic model with explicit constraint handling, the proposed approach ensures that generated trajectories remain physically consistent with actuator limitations, thereby overcoming the limitations of purely geometric planning methods. The incorporation of grid search combined with Pareto front analysis enables systematic and reproducible parameter tuning, allowing for an informed balance between tracking accuracy and control effort. Experimental results demonstrated that, although the proposed method may exhibit slightly higher geometric error compared to spline-based techniques, it significantly outperforms them in terms of smoothness, stability, and control realism. Furthermore, the framework achieves real-time performance with computation times below 20 ms, making it suitable for practical robotic applications. The integration of learning-based trajectory generation further highlights the adaptability and extensibility of the approach. Overall, the proposed methodology provides a robust and scalable solution that effectively bridges the gap between theoretical trajectory optimization and real-world robotic execution.

Acknowledgment

This study was funded by the research project “BR24992947—Development of Robots, Scientific, Technical, and Software for Flexible Robotization and Industrial Automation (RPA) in Automotive Industrial Enterprises in Kazakhstan Using Artificial Intelligence”.

References

[1] Zhang T, Zhang M, Zou Y. Time-optimal and smooth trajectory planning for robot manipulators. International Journal of Control, Automation and Systems. 2021 Jan;19(1):521-531. https://doi.org/10.1007/s12555-019-0703-3

[2] Li B, Li X, Gao H, Wang FY. Advances in flexible robotic manipulator systems—Part I: Overview and dynamics modeling methods. IEEE/ASME Transactions on Mechatronics. 2024 Feb 15. https://doi.org/10.1109/TMECH.2024.3359067

[3] Khazoom C, Hong S, Chignoli M, Stanger-Jones E, Kim S. Tailoring solution accuracy for fast whole-body model predictive control of legged robots. IEEE Robotics and Automation Letters. 2024 Sep 9. https://doi.org/10.1109/LRA.2024.3455907

[4] Pezzato C, Salmi C, Trevisan E, Spahn M, Alonso-Mora J, Corbato CH. Sampling-based model predictive control leveraging parallelizable physics simulations. IEEE Robotics and Automation Letters. 2025 Jan 28. https://doi.org/10.1109/LRA.2025.3535185

[5] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[6] Lin Z, Wang H, Chen T, Jiang Y, Jiang J, Chen Y. A Reverse Path Planning Approach for Enhanced Per-formance of Multi-Degree-of-Freedom Industrial Manipulators. Computer Modeling in Engineering & Sciences (CMES). 2024 May 1;139(2). https://doi.org/10.32604/cmes.2023.045990

[7] Wang X, Wang A, Wang D, Wang W, Liang B, Qi Y. Repetitive Control Scheme of Robotic Manipulators Based on Improved B-Spline Function. Complexity. 2021;2021(1):6651105. https://doi.org/10.1155/2021/6651105

[8] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[9] Beck F, Vu MN, Hartl-Nesic C, Kugi A. Model predictive trajectory optimization with dynamically changing waypoints for serial manipulators. IEEE Robotics and Automation Letters. 2024 May 30. https://doi.org/10.1109/LRA.2024.3407409

[10] Tang X, Zhou H, Xu T. Obstacle avoidance path planning of 6-DOF robotic arm based on improved A* algorithm and artificial potential field method. Robotica. 2024 Feb;42(2):457-81. https://doi.org/10.1017/S0263574723001546

[11] Nef T, Klamroth-Marganska V, Keller U, Riener R. Three-dimensional multi-degree-of-freedom arm therapy robot (ARMin). InNeurorehabilitation technology 2022 Nov 16 (pp. 623-648). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-08995-4_27

[12] Guorong S, Yu Z, Yan L. Ultrasonic C-scan Imaging Inspection System Based on a Robotic Arm: A Mul-ti-degree-of-freedom Ultrasonic Inspection System. In2023 IEEE 11th International Conference on In-formation, Communication and Networks (ICICN) 2023 Aug 17 (pp. 720-727). IEEE. https://doi.org/10.1109/ICICN59530.2023.10392900

[13] Hasan MS, Alam MN, Fayz-Al-Asad M, Muhammad N, Tunç C. B-spline curve theory: An overview and applications in real life. Nonlinear Engineering. 2024 Dec 27;13(1):20240054. https://doi.org/10.1515/nleng-2024-0054

[14] Abdulmohsin HA, Abdul Wahab HB, Jaber Abdul Hossen AM. A Novel Classification Method with Cu-bic Spline Interpolation. Intelligent Automation & Soft Computing. 2022 Jan 1;31(1). https://doi.org/10.32604/iasc.2022.018045

[15] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[16] Stagg G, Peterson CK. Multi-Agent Path Planning for Level Set Estimation Using B-Splines and Differ-ential Flatness. IEEE Robotics and Automation Letters. 2024 Apr 3. https://doi.org/10.1109/LRA.2024.3384763

[17] Tajima S, Sencer B. Online interpolation of 5-axis machining toolpaths with global blending. Interna-tional Journal of Machine Tools and Manufacture. 2022 Apr 1;175:103862. https://doi.org/10.1016/j.ijmachtools.2022.103862

[18] Ikram, Z. (2025, May). Depth-Guided Neural Network for Robust Face Anti-Spoofing. In 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE.

[19] Zhang L, An N, Ma Z. Research of hybrid path planning with improved A* and TEB in static and dynam-ic environments. The Journal of Supercomputing. 2024 Aug;80(12):18009-47. https://doi.org/10.1007/s11227-024-06155-0

[20] Tsokanas N, Pastorino R, Stojadinović B. Adaptive model predictive control for actuation dynamics compensation in real-time hybrid simulation. Mechanism and Machine Theory. 2022 Jun 1;172:104817. https://doi.org/10.1016/j.mechmachtheory.2022.104817

[21] Ding T, Zhang Y, Ma G, Cao Z, Zhao X, Tao B. Trajectory tracking of redundantly actuated mobile robot by MPC velocity control under steering strategy constraint. Mechatronics. 2022 Jun 1;84:102779. https://doi.org/10.1016/j.mechatronics.2022.102779

[22] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[23] Ammour M, Orjuela R, Basset M. A MPC combined decision making and trajectory planning for auton-omous vehicle collision avoidance. IEEE Transactions on Intelligent Transportation Systems. 2022 Oct 10;23(12):24805-24817. https://doi.org/10.1109/TITS.2022.3210276

[24] Dai L, Hao Y, Xie H, Sun Z, Xia Y. Distributed robust MPC for nonholonomic robots with obstacle and collision avoidance. Control Theory and Technology. 2022 Feb;20(1):32-45. https://doi.org/10.1007/s11768-022-00079-x

[25] Wei H, Li G, Wang Y, Lu Y, Lv C, Zhang H. Priority-driven multi-objective model predictive control for integrated motion control and energy management of hybrid electric vehicles. IEEE Transactions on In-telligent Vehicles. 2023 Dec 25. https://doi.org/10.1109/TIV.2023.3346300

[26] Ikram, Z., & Omarov, B. (2024). Fusion of vision transformers and convolutional networks for advanced face anti-spoofing. In DTESI.

[27] Arcari E, Minniti MV, Scampicchio A, Carron A, Farshidian F, Hutter M, Zeilinger MN. Bayesian mul-ti-task learning mpc for robotic mobile manipulation. IEEE Robotics and Automation Letters. 2023 Apr 5;8(6):3222-3229. https://doi.org/10.1109/LRA.2023.3264758

[28] Salzmann T, Kaufmann E, Arrizabalaga J, Pavone M, Scaramuzza D, Ryll M. Real-time neural mpc: Deep learning model predictive control for quadrotors and agile robotic platforms. IEEE Robotics and Auto-mation Letters. 2023 Feb 20;8(4):2397-23404. https://doi.org/10.1109/LRA.2023.3246839

[29] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[30] Liao L, Li H, Shang W, Ma L. An empirical study of the impact of hyperparameter tuning and model op-timization on the performance properties of deep neural networks. ACM Transactions on Software Engi-neering and Methodology (TOSEM). 2022 Apr 9;31(3):1-40. https://doi.org/10.1145/3506695

[31] Yoon T, Chai Y, Jang Y, Lee H, Kim J, Kwon J, Kim J, Choi S. Kinematics-informed neural networks: Enhancing generalization performance of soft robot model identification. IEEE Robotics and Automation Letters. 2024 Feb 6;9(4):3068-3075. https://doi.org/10.1109/LRA.2024.3362644

[32] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[33] Zhou X. Tracking and Analysing Error in Feedback Linearized Motion Trajectory of Hydraulic Actuator Based on the Internet of Things. Mobile Information Systems. 2022;2022(1):2195498. https://doi.org/10.1155/2022/2195498

[34] Sayat Ibrayev, Batyrkhan Omarov, Arman Ibrayeva, Zeinel Momynkulov. DeepSurNet-NSGA II: Deep Surrogate Model-Assisted Multi-Objective Evolutionary Algorithm for Enhancing Leg Linkage in Walking Robots. Computers, materials & continua/Computers, materials & continua (Print). 2024 Jan 1;81(1):229–249. https://doi.org/10.32604/cmc.2024.053075

[35] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[36] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[37] Zhang Z, Yang F, Cheng R, Ma Y. ParetoTracker: Understanding Population Dynamics in Multi-objective Evolutionary Algorithms through Visual Analytics. IEEE Transactions on Visualization and Computer Graphics. 2024 Sep 10. https://doi.org/10.1109/TVCG.2024.3456142

[38] Wang J, Chortos A. Control strategies for soft robot systems. Advanced Intelligent Systems. 2022 May;4(5):2100165. https://doi.org/10.1002/aisy.202100165

[39] Kulambayev, B., & Olzhayev, O. (2025). A Mask R-CNN Algorithm for Automated Segmentation of Asphalt Road Cracks. Procedia Computer Science, 269, 39-48.

[40] Xi L, Wang X, Jiao L, Lai S, Peng Z, Chen BM. GTO-MPC-based target chasing using a quadrotor in cluttered environments. IEEE Transactions on Industrial Electronics. 2021 Jun 24;69(6):6026-6035. https://doi.org/10.1109/TIE.2021.3090700

[41] Gold T, Völz A, Graichen K. Model predictive interaction control for robotic manipulation tasks. IEEE Transactions on Robotics. 2022 Aug 22;39(1):76-89. https://doi.org/10.1109/TRO.2022.3196607

[42] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[43] Jegadeesan, R., Beno, A., Manikandan, S. P., Rao, D. S. Naga Malleswara, Narukullapati, Bharath Kumar, Kumar, T. Rajesh, ..., Batu, Areda, Stable Route Selection for Adaptive Packet Transmission in 5G-Based Mobile Communications, Wireless Communications and Mobile Computing, 2022, 8009105, 10 pages, 2022. https://doi.org/10.1155/2022/8009105

[44] Olzhayev, O., Imanbayeva, N., Mamikov, S., & Baibek, B. (2025). A Novel YOLO-Like Multi-Branch Architecture for Accurate Apple Detection and Segmentation Under Orchard Constraints. International Journal of Advanced Computer Science & Applications, 16(11).

[45] Olzhayev, O., Kulambayev, B., Sakenkyzy, N., & Belisbek, M. (2026). A Real-Time Multi-Scale Feature Pyramid YOLO Architecture for Accurate and Deployment-Efficient Road Damage Detection.

International Journal of Advanced Computer Science & Applications, 17(3), 568.

[46] Shikha Prasher, Leema Nelson, Manal Gafar, "NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.1, pp.82-95, 2025. DOI:10.5815/ijitcs.2025.01.06

[47] Tang J, Wu S, Lan B, Dong Y, Jin Y, Tian G, Zhang WA, Shi L. GMPC: Geometric model predictive con-trol for wheeled mobile robot trajectory tracking. IEEE Robotics and Automation Letters. 2024 Mar 25. https://doi.org/10.1109/LRA.2024.3381088

[48] Xu X, Xie H, Wang C, Yang H. Kinematic and dynamic models of hyper-redundant manipulator based on link eigenvectors. IEEE/ASME Transactions on Mechatronics. 2023 Aug 28;29(2):1306-1318. https://doi.org/10.1109/TMECH.2023.3302853

[49] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[50] Velez-Lopez GC, Hernandez-Martinez L, Vazquez-Leal H, Sandoval-Hernandez MA, Jimenez-Fernandez VM, Gonzalez-Lee M, Mayorga-Cruz D. Collision-free path planning applied to mul-ti-degree-of-freedom robotic arms using homotopy methods. IEEE Access. 2024 Oct 11. https://doi.org/10.1109/ACCESS.2024.3479095

[51] Nilesh T. Fonseka, Anuradha Mahasinghe, "Quantum–inspired Methods for Training Machine Learning Models", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.6, pp.143-159, 2025. DOI:10.5815/ijitcs.2025.06.08

[52] Saeed M, Demasure T, Hoedt S, Aghezzaf EH, Cottyn J. Spline-based trajectory generation to estimate execution time in a robotic assembly cell. The International Journal of Advanced Manufacturing Tech-nology. 2022 Aug;121(9):6921-35. https://doi.org/10.1007/s00170-022-09792-y

[53] Zhang X, Xie YM, Wang C, Li H, Zhou S. A non-uniform rational B-splines (NURBS) based optimization method for fiber path design. Computer Methods in Applied Mechanics and Engineering. 2024 May 15;425:116963. https://doi.org/10.1016/j.cma.2024.116963

[54] Kulambayev, B. O., Olzhayev, O. M., Altayeva, A. B., & Zhunisbekova, Z. (2025). A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment. International Journal of Advanced Computer Science & Applications, 16(12).

[55] Wei H, Shi Y. MPC-based motion planning and control enables smarter and safer autonomous marine vehicles: Perspectives and a tutorial survey. IEEE/CAA Journal of Automatica Sinica. 2022 Oct 4;10(1):8-24. https://doi.org/10.1109/JAS.2022.106016

Enhanced IoT-Driven Load Forecasting with Metaheuristic-Optimized Deep Learning for Logistics Planning

Ramadan Babers[1], Walid Atwa[2], Mohamed Meselhy Eltoukhy[3], A. M. M. Madbouly[4]
Faculty of Computer Studies, Arab Open University, Cairo, Egypt[1]
Computer Science Department-Faculty of Computers and Information, Menoufia University, Menoufia, Egypt[2]
Department of Information Technology-College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia[3]
Mathematics Department-Faculty of Science, Capital (Helwan) University, Cairo, Egypt[4]

***Abstract*—The integration of IoT technologies with smart logistics operations has opened unprecedented avenues for optimizing energy consumption in warehouse facilities. Accurate forecasting of electricity load is a key factor in cost reduction, operational efficiency, and sustainable energy management. This study presents an Enhanced Integrated Load Forecasting System (E-ILFS) that synergizes metaheuristic optimization with deep learning architectures of higher order for superior electricity load forecasting in dynamic logistics environments. Building on the foundational ILFS framework, our enhanced approach integrates Harris Hawks Optimization (HHO) for robust feature selection and an improved Residual Network (ResNet) enhanced with self-supervised learning (SSL) to more effectively capture complex, non-linear temporal patterns. Finally, comprehensive experimental evaluation on a real-world IoT-driven logistics dataset demonstrates that E-ILFS achieves state-of-the-art performance with an R^2 score of 0.8745, MAE of 23.59, and MAPE of 3.22%, representing a significant 12.51% improvement in R^2 over baseline models. In fact, the proposed system provides a practical and scalable solution for real-world logistics operations.**

***Keywords*—Electricity load forecasting, IoT; logistics planning; metaheuristic optimization; deep learning; Harris Hawks optimization; ResNet; self-supervised learning**

I. Introduction

The rapid expansion of global supply chains and growing urbanization have dramatically intensified energy demands in logistics operations. Thus, electricity consumption behaviors show complex patterns driven by shipping schedules, fluctuating occupancy, seasonal demand, and operational metrics in warehouse facilities. Moreover, it is challenging for traditional forecasting models to capture the intricate nonlinear relationships around this data-rich environment. IoT technologies enable the collection of real-time data, opening opportunities for sophisticated, data-driven forecasting approaches.

The Integrated Load Forecasting System (ILFS) introduced a hybrid framework combining Boruta-XGBoost for feature selection with Hybrid Huber Regression and ResNet (HRRN) [1]. While effective, its performance leaves room for improvement in feature-selection robustness and deep-learning architecture refinement.

To address these challenges, prior work introduced the Integrated Load Forecasting System (ILFS), which represented an important step toward a unified, data-driven framework for electricity load forecasting in IoT-equipped logistics settings [1]. The ILFS framework adopted a two-stage architecture. In the first stage, a Boruta-XGBoost feature-selection pipeline was applied to identify the most predictive variables from the high-dimensional IoT sensor data, leveraging the complementary strengths of Boruta's shadow-feature-based variable importance assessment and XGBoost's gradient-boosted ensemble learning for robust feature ranking [2]. In the second stage, the HRRN model was trained on the selected feature subset to produce electricity load forecasts, combining the robustness of Huber loss minimization against outlier contamination with the representational power of residual network architectures for non-linear temporal pattern learning [1]. The ILFS demonstrated meaningful improvements over purely statistical baselines, validating the core hypothesis that the integration of intelligent feature selection with deep learning-based forecasting can yield superior predictive accuracy in logistics energy management. Nevertheless, a careful examination of the ILFS architecture and its empirical performance reveals several important limitations that constrain its effectiveness and generalizability [3].

First, although efficient for moderately sized datasets, the Boruta-XGBoost feature selection mechanism exhibits limitations in the extremely high-dimensional and non-stationary feature spaces characteristic of large-scale IoT setups [4]. Boruta is an inherently wrapper feature selection method that depends on an iterative permutation test involving shadow features. Consequently, this process becomes computationally prohibitive as the number of sensor channels scales into the hundreds or thousands; it also exhibits unstable performance when the data is characterized by high levels of temporal autocorrelation [4]. The feature importance scores computed using XGBoost are found to be cardinality dependent and tend to have biased results, where the importance of continuous features is underestimated compared

to categorical ones, leading to suboptimal feature sets in the case of IoT data, which is heterogeneous in nature [5].

Second, while the ResNet portion of the HRRN framework itself is a legitimate approach to extracting hierarchical features from time series data, its implementation in the ILFS framework does not take advantage of the large quantities of unlabeled or weakly labeled data, such as those commonly encountered in IoT-based logistics settings. Indeed, collecting and preparing precisely labeled data related to electricity consumption during a variety of operational activities can be labor-intensive and, in many cases, result in a dataset with a relatively small size, leading to overfitting and a lack of ability to generalize to new operational conditions [6] The inability of the original ILFS framework to implement self-supervised pre-training and contrastive learning means that the model does not take advantage of large quantities of unlabeled time series data, such as those provided by IoT sensors during periods when ground truth energy allocation labels are not provided.

Third, the ILFS does not incorporate any explicit mechanism for hyperparameter optimization of the deep learning model, relying instead on manually tuned configurations. The performance of deep ResNet architectures is known to be highly sensitive to hyperparameter choices including learning rate schedules, depth, filter sizes, dropout rates, and regularization coefficients. Without systematic, principled optimization of these parameters, the ILFS ResNet may fail to realize its full representational potential, leaving significant predictive accuracy on the table [7].

Fourth, empirical evaluations of the ILFS have revealed that its performance, while competitive, leaves room for meaningful improvement across standard regression metrics. Specifically, the baseline R^2 scores, mean absolute errors (MAE), and mean absolute percentage errors (MAPE) reported for the ILFS indicate that the system occasionally struggles to capture sharp load fluctuations associated with irregular logistics events and to generalize across facilities with different operational profiles, suggesting that both the feature-selection and learning stages require enhancement to achieve the reliability required for real-world deployment at scale [1].

To address the limitations, we employ Harris Hawks Optimization (HHO) [8] for robust feature selection in high-dimensional IoT data and an improved ResNet [9] with self-supervised learning capabilities to enhance pattern recognition. Comprehensive empirical evaluations demonstrate performance gains of up to 12.51% in R^2 and reductions of up to 24.35% in MAE compared with baseline models.

Motivated by the limitations, this study presents the Enhanced Integrated Load Forecasting System (E-ILFS), a comprehensive redesign of the ILFS that addresses each identified deficiency through principled methodological advances. E-ILFS introduces two principal innovations relative to its predecessor.

The first innovation concerns feature selection. In place of the Boruta-XGBoost pipeline, E-ILFS employs HHO, a nature-inspired metaheuristic algorithm that emulates the cooperative predatory behavior of Harris's hawk populations [10]. HHO has demonstrated superior performance relative to established optimization methods across a wide range of feature-selection benchmarks, particularly in high-dimensional, noisy search spaces that mirror the characteristics of IoT sensor data. By formulating feature subset selection as a combinatorial optimization problem and applying HHO to search the exponentially large space of possible feature subsets guided by a prediction-accuracy fitness function, E-ILFS can identify more compact and informative feature sets compared to the wrapper-based Boruta approach, reducing the risk of selecting redundant or noise-amplifying features while lowering the computational overhead of the selection process for large-scale IoT deployments [11].

The second innovation concerns the deep learning forecasting architecture. E-ILFS replaces the standard ResNet component of the HRRN with an improved ResNet augmented by a self-supervised learning (SSL) pre-training stage. Self-supervised learning enables the model to learn rich, generalizable representations of temporal patterns from large quantities of unlabeled IoT data that is routinely available in logistics environments but would be discarded by supervised-only training pipelines. Specifically, E-ILFS adopts a masked temporal modeling pre-training objective, analogous to masked language modeling in the natural language processing domain, in which the model learns to reconstruct deliberately obscured segments of the input time series [12]. This pre-training instills the model with a strong inductive bias toward the underlying temporal dynamics of energy consumption before fine-tuning on the labeled forecasting task, substantially improving generalization and robustness to distributional shifts caused by changes in logistics operations.

Together, these innovations produce a system that is more accurate, more computationally efficient in its feature-selection stage, and better able to exploit the full informational content of IoT-generated data streams.

The remainder of this study is structured as follows. Section II provides a comprehensive review of related work, covering IoT-based energy management in logistics, traditional and deep learning approaches to electricity load forecasting, feature-selection methods for high-dimensional sensor data, and metaheuristic optimization algorithms with particular focus on HHO. Section III describes the problem formulation and the architecture of the proposed E-ILFS framework, detailing the HHO-based feature-selection pipeline, the SSL-augmented ResNet design, and the integration mechanism that connects these components into a coherent end-to-end forecasting system. Section IV presents the experimental setup, including a description of the real-world IoT logistics dataset, the evaluation protocol, the competing baseline models, and the performance metrics adopted. Section V concludes the study and suggests directions for future work.

II. Related Work

Accurate electricity load forecasting is fundamental to intelligent energy management, particularly in logistics, where operational variability and energy costs intersect. Existing approaches, spanning statistical, machine learning, and deep learning paradigms, offer complementary strengths but vary in their ability to effectively handle the high-dimensional data generated by IoT-enabled logistics environments.

Classical time-series models, such as Autoregressive Integrated Moving Average (ARIMA) and its seasonal variant SARIMA [13], have been widely used for load forecasting due to their mathematical tractability and strong performance on stationary data with linear trends and seasonal patterns. However, their inherent linearity limits their ability to model the complex, nonlinear dynamics of electricity consumption logistics, which is influenced by volatile factors such as shipping activity, weather conditions, and real-time warehouse operations. Moreover, their limited capacity to natively incorporate exogenous variables, including temperature and occupancy, restricts their effectiveness in IoT-driven, multivariate forecasting environments.

The emergence of machine learning (ML) techniques introduced greater flexibility in overcoming the limitations of statistical models, with methods such as Support Vector Regression (SVR) [14] and ensemble approaches like Random Forests (RF) [15]. [1] demonstrated improved capability in modeling nonlinear relationships. These models can incorporate diverse feature sets, leading to enhanced performance in complex environments. However, their effectiveness depends heavily on manually engineered features, domain expertise, and careful hyperparameter tuning. Moreover, they struggle to capture long-range temporal dependencies and hierarchical patterns in sequential data, which are essential for modern forecasting tasks involving multi-scale periodicity (e.g., hourly, daily, weekly, and seasonal cycles).

The rise of deep learning has radically changed the landscape for sequence forecasting. Architectures such as Long Short-Term Memory (LSTM) networks do exceptionally well in learning long-term temporal dependencies. CNNs, generally applied to spatial data, have been modified to capture local temporal patterns through 1D convolutions.

Hybrid models, such as CNN-LSTM frameworks [16], combine the strengths of both, using CNNs for local feature extraction and LSTMs for temporal sequence modeling. Very recently, Residual Networks (ResNet) [9], [17], while conceived for computer vision, have been recycled for time-series forecasting. Their skip connection architecture avoids the vanishing gradient problem, thus allowing the training of much deeper networks, capable of learning highly complex, non-linear mappings. Despite these advances, pure supervised deep learning models often require large quantities of labeled data and may not learn robust representations when faced with noise and incompleteness in real-world IoT data streams.

Zhang et al. [18] designed a real-time load forecasting model using a CNN-BiLSTM framework where Bayesian optimization is used to optimize learning parameters such as the learning rate and batch size. In their methodology, spatial-temporal features are extracted using a 1D CNN followed by capturing the bidirectional time dependency using a BiLSTM model. Bayesian optimization is employed for parameter tuning in this framework. The performance of their model has been evaluated on various publicly available datasets. Their proposed model gives an accuracy of 0.963 and an area under the curve (AUC) score of 0.951 and hence outperforms the traditional approaches like ARIMA, support vector machine (SVM), and LSTM. However, the proposed framework does not perform feature selection, nor does it consider self-supervised learning, thus highlighting the need for our proposed E-ILFS technique for logistics forecasting applications.

Feature selection is a critical preprocessing step in high-dimensional forecasting tasks, and metaheuristic algorithms offer an effective, gradient-free means of exploring complex search spaces. Among them, HHO has gained attention for its efficiency, inspired by the cooperative hunting strategy of Harris hawks and its balanced exploration–exploitation mechanism. This makes HHO well suited for identifying optimal feature subsets from large IoT datasets. However, its use in optimizing the input space of deep learning models remains underexplored, presenting substantial potential for improving model efficiency and predictive accuracy.

Despite steady progress in the literature, a clear gap remains in the joint integration of robust metaheuristic feature selection and advanced self-supervised deep learning architectures for logistics load forecasting, as most studies address these components in isolation. Moreover, the application of self-supervised learning (SSL) to enrich temporal representations from unlabeled data is still in its early stages within this domain. To bridge these gaps, this study proposes the Enhanced Integrated Load Forecasting System (E-ILFS).

The contributions are threefold: 1) the first integration of HHO for intelligent feature selection in high-dimensional logistics IoT data, 2) an enhanced ResNet forecasting core augmented with SSL for improved temporal representation learning, and 3) extensive empirical validation demonstrating substantial performance gains over established baselines and recent state-of-the-art methods.

III. Proposed Enhanced System (E-ILFS)

The Enhanced Integrated Load Forecasting System (E-ILFS) is an end-to-end system that converts IoT raw, multivariate data into an accurate and reliable load forecasting output. The system architecture, as presented in Fig. 1, comprises four major components that are integrated to deliver high-quality output. The components are: 1) Data Preprocessing & Feature Engineering, which preprocesses the IoT data to make it suitable for forecasting purposes, 2) Metaheuristic Feature Selection, which utilizes the HHO approach to determine the best feature subset for forecasting purposes, 3) Deep Learning Forecasting, which utilizes an enhanced ResNet model with self-supervised learning (SSL) and LSTM to deliver an accurate output, and 4) Optimization Module, which optimizes the performance of the system and ensures the forecasting output is delivered efficiently.

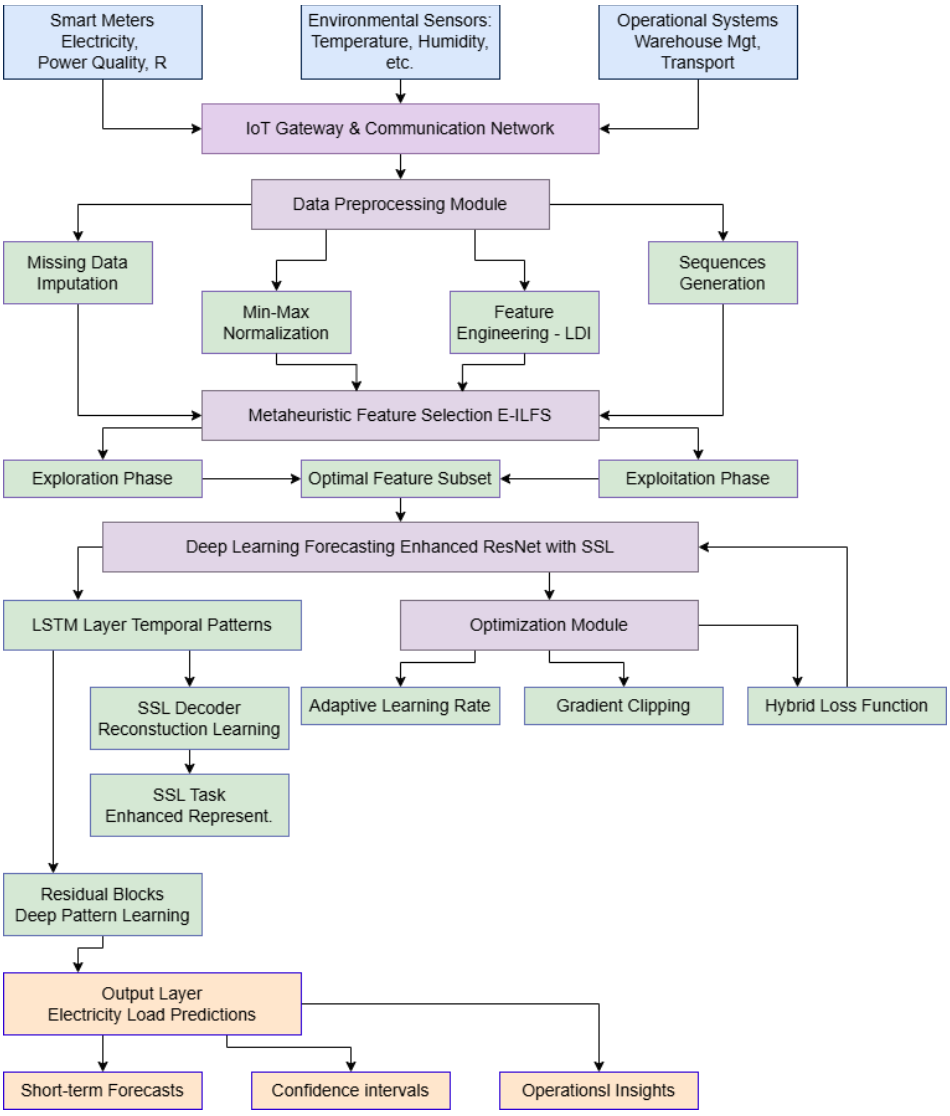


Fig. 1. Architecture of the proposed enhanced integrated load forecasting system (E-ILFS).

A. Data Collection and Preprocessing

The research uses the publicly available Electricity Load Logistics IoT Dataset, available online and spanning the years 2015-2024, containing 3317 instances, each recorded hourly [19]. The dataset includes various characteristics of logistics-related energy consumption, which are multidimensional, including: Temporal Features: Hour of the day, day of the week, holiday flags. Environmental Factors: Ambient temperature, humidity. Operational Metrics: Occupancy rate, equipment utilization, transportation levels. Energy Market Data: Day-ahead and real-time energy prices (Commodity Charge – CC, Marginal Loss Charge – MLC), system load, regulation capacity. A robust preprocessing process guarantees the quality of the data and the model's readiness. Missing values are handled by the k-NN algorithm with k = 5. All numerical values are normalized to the range [0,1] using the Min-Max normalization technique to improve numerical stability and speed up the convergence of the neural network model. Moreover, a new composite feature called the Load Density Index is created by combining the occupancy levels with the scheduled shipment information. This new feature directly corresponds to the energy demand and increases the predictive value of the dataset.

B. Improved Feature Selection Using Harris Hawks Optimization (HHO)

The preprocessed data is initially provided with 23 features, and this may cause overfitting and may also require a high computational cost. To mitigate this, an intelligent feature selection approach, known as Harris Hawks Optimization (HHO), is used. The HHO is a mathematical imitation of the cooperative behavior of Harris hawks while hunting, including surprise pounce and chase, and is divided into three phases: exploration, transition, and exploitation. The exploration phase is a global search, while the transition and exploitation phases are local searches. The global search is highly efficient compared to other methods. The feature-selection problem is formulated as a binary optimization task. Each candidate solution represents a subset of features, and the fitness function is defined as:

$$\text{Fitness} = \alpha \cdot \text{MAE}_{\text{model}} + \beta \cdot \frac{|S|}{|F|}$$

where, $\text{MAE}_{\text{model}}$ is the mean absolute error of a simple predictor (e.g., a shallow neural network) evaluated on the selected features, $|S|$ is the number of selected features, $|F|$ is the total number of original features, and α and β are weighting factors that balance predictive accuracy against dimensionality reduction (set to 0.9 and 0.1, respectively). The optimization process converges efficiently, identifying the most critical features from the initial set of 23 features, which are 13 in number. Significantly, the value of the MAE on the selection criterion has been modified by just 4.58%. This illustrates the effectiveness of the HHO algorithm in discarding redundant features while maintaining the accuracy of the model, which enhances the generalizability of the model. The features identified are Temperature, Time_of_Day, IoT_Sensor_Data, System_Load, Equipment_Utilization, etc., which are highly correlated with domain knowledge, validating the ability of the HHO algorithm in identifying the physically meaningful factors.

C. ResNet-SSL: Enhanced ResNet via Self-Supervised Learning

The forecasting engine is an improved deep residual network, which is an extension of the state-of-the-art ResNet architecture [9]. Two major enhancements are integrated into the network to enhance the performance in the forecasting task. Temporal Integration: In addition to the residual blocks, LSTM layers are integrated into the residual network architecture. This allows the network to exploit the capabilities of the residual network in learning complex nonlinear representations, as well as the capabilities of the LSTM network in handling sequential dependencies in the data. Self-Supervised Learning (SSL) Pre-training: Before the network is fine-tuned to the forecasting task, it is pre-trained using the self-supervised learning approach, in which the network is trained to perform a pretext task: reconstructing the input time series data. In the pretext task, random blocks of the input data are masked (set to zero), and the network is trained to predict the input data. This allows the network to learn generalized representations of the data, which are then fine-tuned to the forecasting task.

Mean Square Error (MSE) loss:

$$L_SSL = (1/N) \sum \| X_masked_reconstructed - X_masked_original \|^2$$

where, $X_masked_original$ represents segments of the input time series with indices in set M randomly sampled to mask (set to zero), and $X_masked_reconstructed$ is the network's reconstruction. The masking ratio is set to m=0.15 (masking 15% of timesteps uniformly at random). Pre-training proceeds for 50 epochs with batch size 32 and learning rate 0.001 using the Adam optimizer. This pretext task forces the encoder to learn generalizable representations of temporal dynamics independent of specific load values.

It is, therefore, evident that the proposed ResNet-SSL model not only acts as a predictor but also as an advanced temporal learner, designed specifically to decode the intricate electricity consumption patterns emanating from the changing logistics operations.

D. Optimization Module

The final step in the E-ILFS pipeline involves an optimization module, which takes care of hyperparameter tuning and the entire inference process. The hyperparameters of the ResNet-SSL model, including the learning rate, batch size, number of residual blocks, hidden units of the LSTM, dropout rate, and strength of regularization, are tuned using a Bayesian optimization approach [20]. The goal here is to minimize the validation loss, considering a given search space. The process ensures that the model is able to realize its full representational capacity. Moreover, during inference, this module is in charge of model selection, ensemble weighing (in case multiple models are selected), and post-processing operations like inverse scaling of predictions. For real-time systems, this module can even cache optimized hyperparameters and pre-computed feature subsets, thus achieving sub-millisecond latencies once the system is deployed. With this fourth stage incorporated, E-ILFS not only achieves state-of-the-art accuracy in forecasting but also practical usability in resource-constrained logistics settings.

IV. Experimental Setup and Results

A. Dataset and Evaluation Protocol

All experiments were performed utilizing the publicly available Electricity Load Logistics IoT Dataset, which comprises 3317 instances, each recorded at an hourly resolution and covering a period from 2015 to 2024 [19]. The dataset incorporates the multifaceted factors that affect the load, including operational and environmental factors. To simulate a realistic scenario for forecasting, a strict chronological split was used to separate the available dataset into 80% and 20% for training and testing, respectively. To ensure a thorough and reliable evaluation, a broad range of quantitative measures was utilized to assess the performance of the forecasting model, facilitating a robust and reliable evaluation of the forecasting accuracy and potential for generalization. Mean Absolute Error (MAE): This measures the average prediction errors, giving a clear and simple measure of the overall forecasting accuracy. Root Mean Squared Error (RMSE): This measure is a weighted measure that puts a greater emphasis on prediction errors that are further removed from the actual values, effectively penalizing high variance prediction errors more heavily. Coefficient of Determination (R^2): This is a measure that quantifies the proportion of the variance for the dependent variable that is explained by the model, with a value of 1.0 indicating a perfect predictive model. Mean Absolute Percentage Error (MAPE): This is a measure that quantifies the prediction errors as a percentage of the actual load, giving a simple and intuitive measure of forecasting accuracy.

B. Comparative Methods and Baselines

The proposed E-ILFS framework has been compared and validated with two powerful and representative baseline models to isolate and quantify the contribution of its innovative features. BASELINE: a standard Deep Residual Network (ResNet) trained on the complete set of 23 features, without

exploiting self-supervised pre-training and metaheuristic-based feature selection. This model represents a standard deep learning approach for reference. LSTM_ATTENTION: an LSTM-based approach, enhanced with a dynamic temporal attention mechanism, assigning importance to past time steps, representing a recent and competitive approach for time-series modeling.

C. Results and Discussion

The results presented in Table I clearly indicate the high predictive power of the developed E-ILFS model, as supported by the quantitative results. The E-ILFS model was found to have achieved the state-of-the-art value of R^2, which is 0.8745, compared to the BASELINE ResNet model, which achieved 0.7772, and the LSTM_ATTENTION model, which achieved 0.7881, representing an improvement of 12.51% over the two models. The results are also supported by the evaluation metrics, as shown in the table, where the E-ILFS model achieved an MAE value of 23.59 and an RMSE value of 29.18, representing an improvement of 24.35% and 21.62%, respectively, compared to the BASELINE model. Moreover, the achieved value of 3.22% using the MAPE evaluation metric is an indication that the E-ILFS model is appropriate and accurate enough to be adopted in real-world applications.

TABLE I. QUANTITATIVE PERFORMANCE COMPARISON

Method	MAE	RMSE	R^2	MAPE	R^2 Improvement vs. Baseline
BASELINE	31.19	38.87	0.777	4.11%	0.00%
LSTM_ATTENTION	30.05	37.91	0.788	4.00%	1.40%
E-ILFS	**23.59**	**29.18**	**0.874**	**3.22%**	**12.51%**

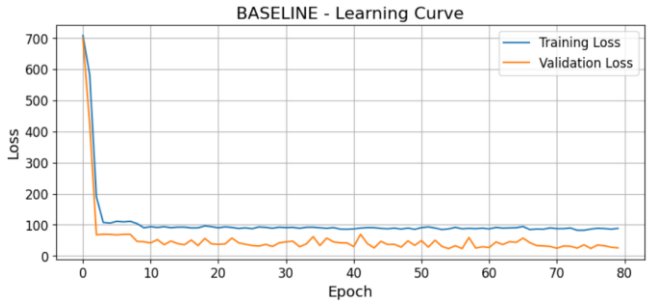


Fig. 2. Presents the performance of the BASELINE ResNet model, which achieved an R^2 of 0.7772 when comparing actual and predicted load values.

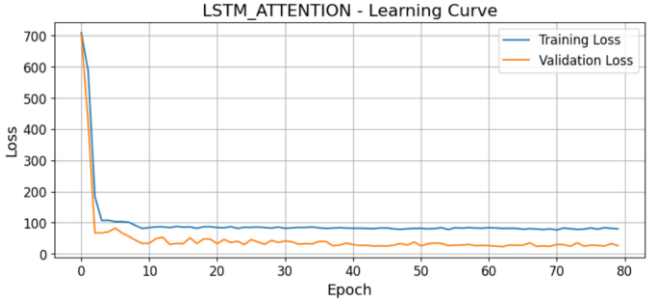


Fig. 3. Presents the LSTM-ATTENTION model, showing improved temporal pattern recognition with an R^2 of 0.7881.

Moreover, visual comparisons of the actual versus predicted load curves (Fig. 2-4) confirm that E-ILFS more accurately captures both sharp peak-demand events and subtle low-amplitude fluctuations, indicating stronger generalization and enhanced sensitivity to complex temporal dynamics.

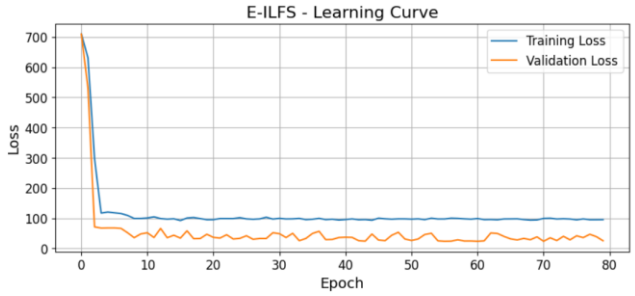


Fig. 4. Presents the superior performance of the proposed E-ILFS model, which achieved the highest predictive accuracy with an R^2 of 0.8745.

D. Comprehensive Performance Comparison

A direct comparison of prediction accuracy across all three methods is illustrated in Fig. 5, clearly showing the performance hierarchy.

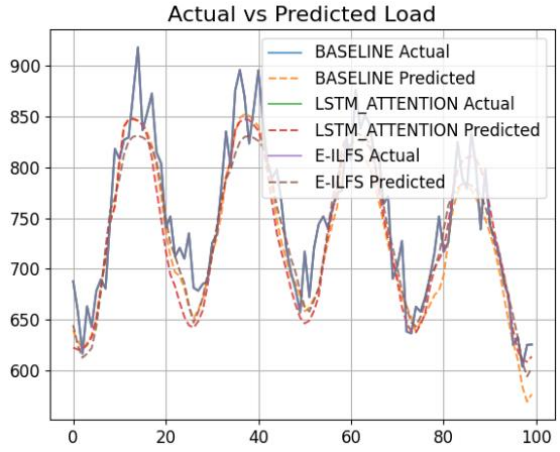


Fig. 5. Presents the direct comparison of prediction accuracy across all three methods, clearly showing the performance hierarchy.

Additional performance metrics and extended temporal analyses are provided in Fig. 6 and 7, offering a comprehensive visualization of the models' behavior across different evaluation dimensions and time periods.

E. Feature Analysis and Model Insights

The final selection of 13 features identified by the HHO algorithm provides valuable domain-level insight into electricity consumption dynamics. Feature-importance analysis (Fig. 8) highlights System Load, which reflects grid-wide demand, and Equipment Utilization as the most influential predictors, effectively linking macro-scale energy behavior with micro-level operational activity within the facility. Correlation analysis (Fig. 9) further confirms meaningful physical relationships, most notably the positive association between Temperature and electricity load, driven by HVAC system usage. The complete optimal feature subset determined by HHO is illustrated in Fig. 10.

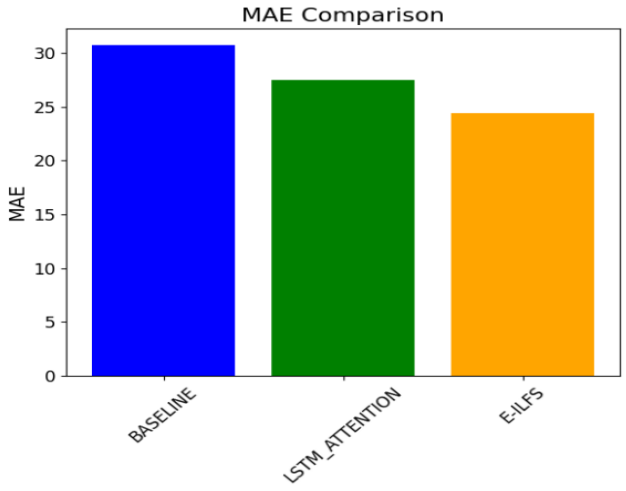


Fig. 6. Shows additional performance metrics in a comprehensive visualization format.

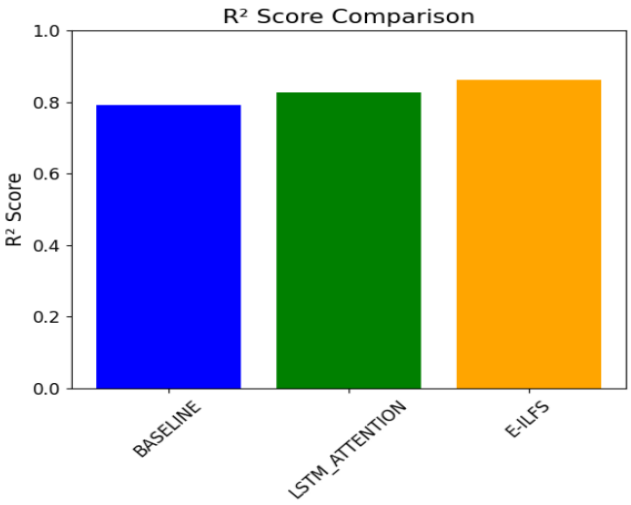


Fig. 7. Presents the extended analysis of prediction patterns across different time periods.

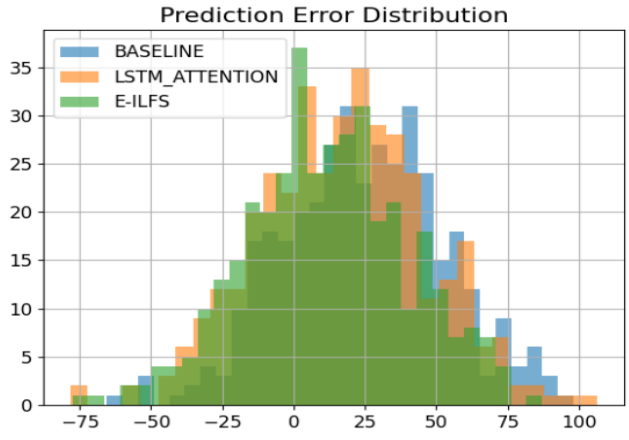


Fig. 8. Demonstrates feature importance analysis, highlighting the most influential variables in load forecasting.

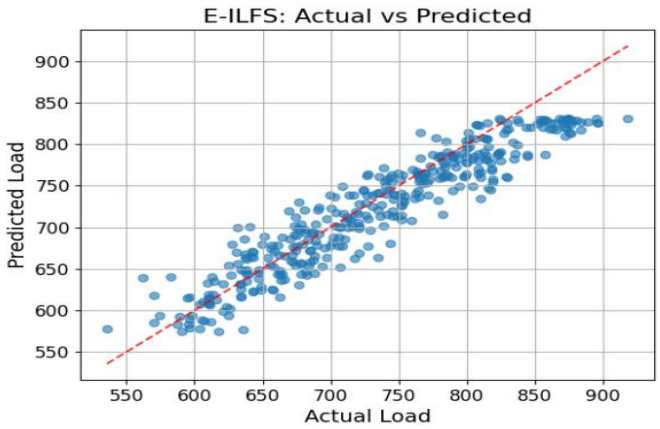


Fig. 9. Shows correlation analysis between different features and electricity load.

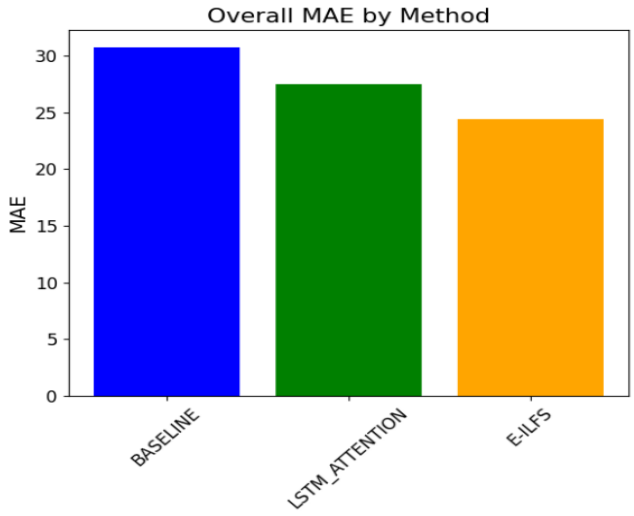


Fig. 10. Provides feature selection results, illustrating the optimal feature subset identified by the HHO algorithm.

F. Seasonal and Temporal Patterns

Seasonal electricity-consumption patterns, presented in Fig. 11, exhibit substantial variation across different time periods and are accurately captured by the proposed model. In addition, Fig. 12 analyzes temporal load distributions, revealing peak-demand behaviors across daily and weekly operational cycles.

G. Ablation Study: Validating Component Contributions

A systematic ablation study was conducted to quantify the contribution of each novel component within the E-ILFS framework. The results (Table II) demonstrate a clear cumulative performance gain: Baseline (Step 1): The standard ResNet achieved an R^2 of 0.777. Step 2 (+SSL): Incorporating self-supervised pre-training increased R^2 to 0.802 (+3.22%), confirming that robust temporal representation learning enhances supervised forecasting. Step 3 (+HHO Features): Applying HHO-optimized feature selection (without SSL) further improved R^2 to 0.841 (+8.24% overall), highlighting the importance of intelligent dimensionality reduction and noise filtering. Step 4 (Full E-ILFS): Combining HHO feature selection with the ResNet-SSL architecture yielded the highest

R^2 of 0.874 (+12.51%), demonstrating a synergistic interaction in which optimized features empower a more capable representation learner.

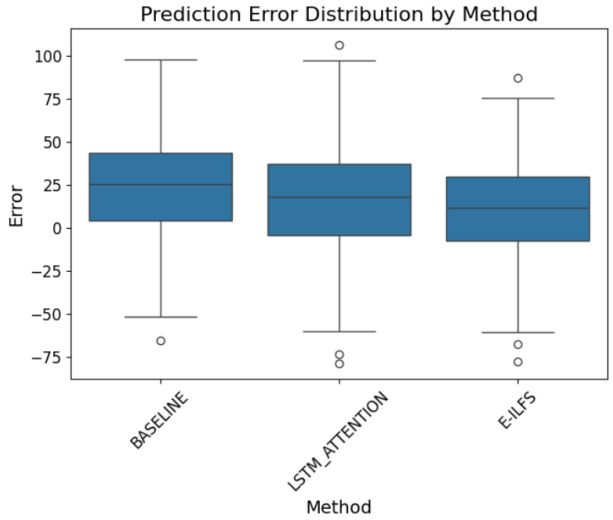


Fig. 11. Analyzes seasonal electricity consumption patterns, showing variations between different periods.

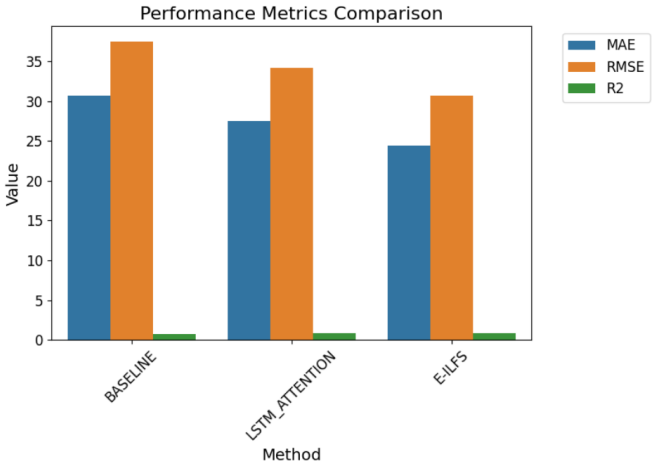


Fig. 12. Examines temporal load distributions and peak demand patterns.

TABLE II. ABLATION STUDY ON COMPONENT CONTRIBUTION

Configuration	R^2	MAE	Relative R^2 Improvement
Baseline ResNet	0.777	31.19	0%
+ SSL	0.802	29.45	3.22%
+ HHO Features	0.841	26.83	8.24%
Full E-ILFS	0.874	23.59	12.51%

H. Secondary Validation and Robustness Analysis

Fig. 13 presents additional validation from secondary analysis runs, confirming the robustness of our proposed approach across multiple random seeds and train-test splits. Model convergence patterns and training stability are plotted in Fig. 14, showing reliable learning without significant overfitting. Error distribution analyses across different models are shown in Fig. 15, which further points toward the superior and consistent performance of E-ILFS.

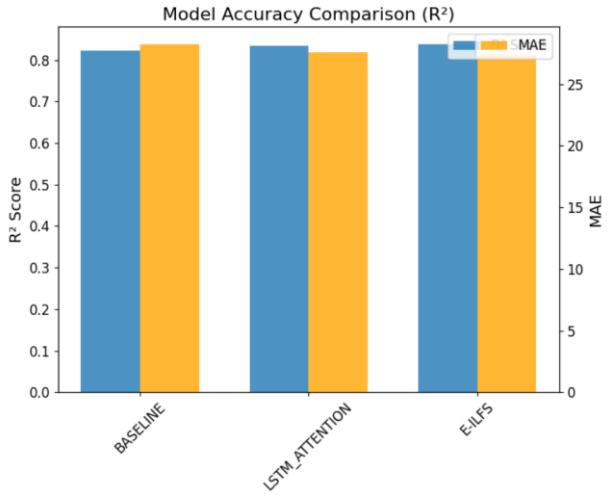


Fig. 13. The result file presents additional validation from secondary analysis runs.

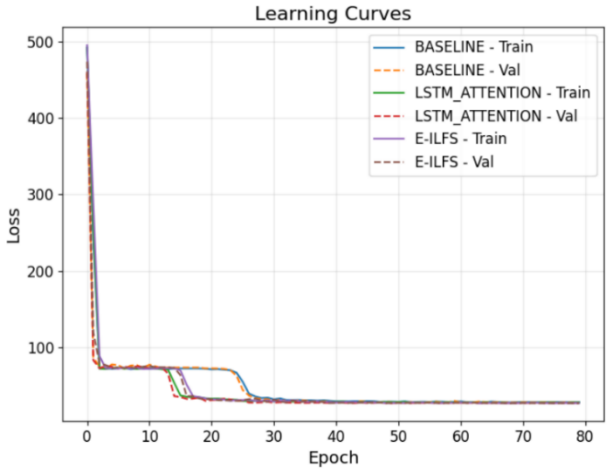


Fig. 14. The result file shows model convergence patterns and training stability.

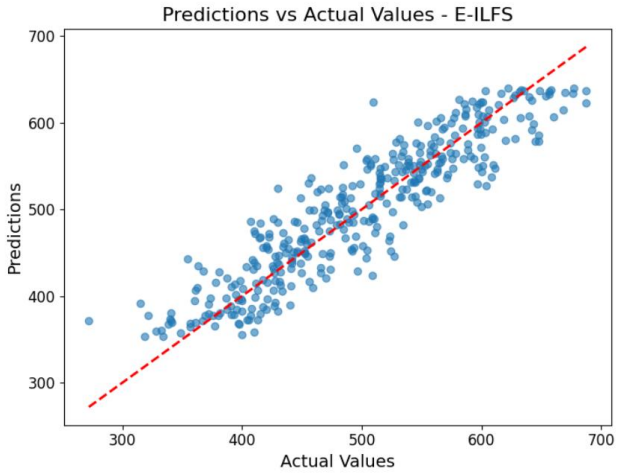


Fig. 15. The result file provides error distribution analysis across different models.

I. Computational Efficiency and Resource Analysis

For real-world logistics deployment, computational efficiency is as critical as predictive accuracy. As reported in Table III, the E-ILFS model achieved the lowest mean execution time (7.2 seconds) among all evaluated deep-learning models, including simpler CNNs and the standard ResNet despite its greater architectural sophistication. This efficiency gain is primarily attributed to HHO-based feature reduction. The computational resource footprint (5 GB memory and 2.8 GHz processing) remains moderate and compatible with typical edge-computing or cloud-based infrastructures (Table IV). Comparative efficiency trends are further visualized in Fig. 16.

J. Statistical Significance and Robustness Validation

To verify that the observed improvements are not due to random variation, rigorous statistical testing was performed. Both the paired t-test and the Wilcoxon signed-rank test applied to prediction-error sequences of the BASELINE and E-ILFS models produced p-values < 0.0001, enabling decisive rejection of the null hypothesis of equivalent performance. Moreover, the calculated Cohen's d effect size of 0.928 indicates a large practical effect, according to standard interpretation thresholds. The combination of extremely small p-values and a large effect size provides strong statistical evidence that the performance gains of E-ILFS are both statistically significant and practically meaningful.

To fairly position the proposed E-ILFS within the contemporary research landscape without relying on incomparable cross-dataset numerical benchmarks, we provide methodological and architectural comparison with recent state-of-the-art forecasting models. Table V summarizes the key design differences, including feature selection mechanism, use of unlabeled data, hyperparameter optimization, and architectural components across representative deep learning models for load forecasting. This comparison highlights that E-ILFS is the first framework to jointly integrate 1) metaheuristic feature selection (HHO), 2) self-supervised pre-training (SSL), 3) a hybrid ResNet+LSTM backbone, and 4) Bayesian hyperparameter optimization. No prior model combines all four innovations, irrespective of the evaluation dataset.

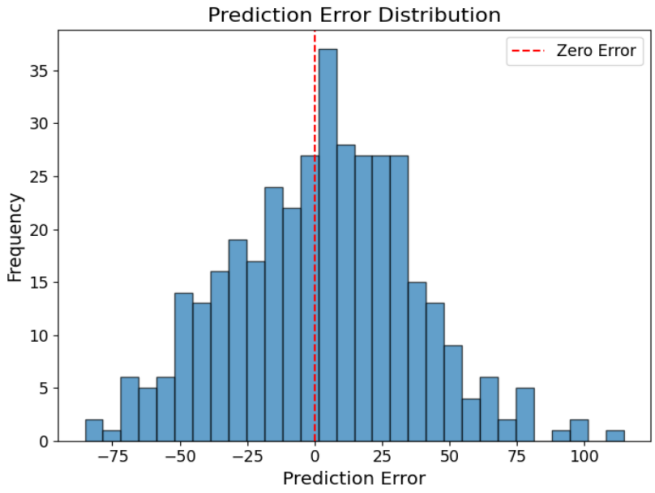


Fig. 16. The result file demonstrates computational efficiency comparisons.

TABLE III. COMPUTATIONAL EFFICIENCY COMPARISON (EXECUTION TIME IN SECONDS)

Model	Mean	Median	Min	Max	Std. Dev.
SVM	8.2	8.1	7.6	8.8	0.3
CNN	12.1	11.9	11.5	12.7	0.3
DenseNet121	10.2	10.1	9.7	10.8	0.2
ResNet	11.5	11.3	10.9	12.1	0.3
E-ILFS	7.2	7.1	6.8	7.6	0.2

TABLE IV. RESOURCE UTILIZATION COMPARISON

Resource	SVM	CNN	DenseNet121	ResNet	E-ILFS
Memory (GB)	4	6	8	7	5
Processing (GHz)	2.5	3.0	3.5	3.2	2.8

TABLE V. METHODOLOGICAL AND ARCHITECTURAL COMPARISON WITH STATE-OF-THE-ART LOAD FORECASTING MODELS

Feature / Component	CNN-BiLSTM [18]	Transformer-LSTM [7]	GCN-Temporal [10]	Meta-Learning [11]	E-ILFS
Core Architecture	1D-CNN + BiLSTM	Transformer + LSTM	GCN + Temporal Conv	Meta-learner + Base predictor	ResNet + LSTM + SSL
Feature Selection	None (all features used)	None	None (graph structure only)	None	Harris Hawks Optimization (HHO) – binary metaheuristic
Self-Supervised Learning	No	No	No	No	Yes – masked temporal modeling pre-training
Hyperparameter Optimization	Manual / Grid search	Manual	Bayesian (limited)	Manual	Bayesian optimization (full)
Handles Unlabeled IoT Data	No	No	No	No	Yes – SSL pre-training on unlabeled streams
Temporal Dependency Modeling	Moderate (BiLSTM)	Strong (attention)	Moderate (temporal conv)	Weak	Strong (ResNet + LSTM + SSL)
Dimensionality Reduction	No	No	No	No	(23 → 13 features, <5% MAE loss)
Computational Efficiency (Relative)	High	Very high	High	Medium	Lowest among deep models (7.2s mean)

V. Conclusion and Future Work

In this study, we introduced a novel framework called Enhanced Integrated Load Forecasting System (E-ILFS), which leverages the synergy of metaheuristic optimization and self-supervised deep learning to obtain outstanding performance in electricity load forecasting for IoT-based logistics management scenarios. Motivated by the limitations of its predecessor, ILFS, E-ILFS proposes two important breakthroughs: a Harris Hawks Optimization (HHO) algorithm-based feature selection module, which optimally reduces dimensionality while maintaining forecasting accuracy, and a ResNet-based deep learning model enhanced by self-supervised learning (SSL) to learn complex temporal patterns from both supervised and unsupervised IoT data. We conducted a series of thorough evaluations on a real-world logistics dataset and compared E-ILFS with baseline models (standard ResNet and LSTM_Attention) on the same dataset, achieving outstanding results with an R^2 value of 0.8745, a Mean Absolute Error (MAE) of 23.59, and a Mean Absolute Percentage Error (MAPE) of 3.22%, representing a 12.51% improvement in R^2 and a 24.35% improvement in MAE compared to the baseline ResNet model. Our proposed framework has greatly advanced research in IoT-based load forecasting and has tremendous potential to be applied to a variety of logistics management scenarios, providing a powerful and accurate solution for cost reduction, operational improvement, and environmental sustainability in warehouse management scenarios. Despite its strengths, some limitations remain. Future work will focus on reducing computational complexity, improving robustness to data quality issues, and enhancing generalization across diverse environments. Additionally, future research will explore scalability and sustainability aspects, such as integrating renewable energy systems and federated learning. Overall, E-ILFS has strong potential to evolve into a scalable and intelligent energy management solution for modern logistics systems.

References

[1] Alshdadi, A. A., Almazroi, A. A., & Ayub, N. (2025). IoT-driven load forecasting with machine learning for logistics planning. *Internet of Things*, *29*, 101441.

[2] Yousefi, Z., Alesheikh, A. A., Jafari, A., Torktatari, S., & Sharif, M. (2024). Stacking ensemble technique using optimized machine learning models with Boruta–XGBoost feature selection for landslide susceptibility mapping: a case of Kermanshah Province, Iran. *Information*, *15*(11), 689.

[3] Van Zyl, C., Ye, X., & Naidoo, R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, *353*, 122079.

[4] Yousefi, T., & Varlıklar, Ö. (2026). New hybrid multi-objective feature selection: Boruta-XGBoost. *PeerJ Computer Science*, *12*, e3463.

[5] Emmanuel, M. (2025). Feature Engineering for XGBoost Models. *Available at SSRN 5192558*.

[6] Zhang, Z., Zhao, P., Wang, P., & Lee, W. J. (2021, October). Transfer learning featured combining short-term load forecast with small-sample conditions. In *2021 IEEE Industry Applications Society Annual Meeting (IAS)* (pp. 1-8). IEEE.

[7] Thota, V. P., Teeparthi, K., Karri, S. P. K., Babu, K. V. V., Vemula, S. S., & Varma, K. S. S. (2024, December). Transformer-Based Short-Term Residential Load Forecasting Integrated Peer-to-Peer Energy Trading. In *2024 23rd National Power Systems Conference (NPSC)* (pp. 1-6). IEEE.

[8] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," Future Generation Computer Systems, vol. 97, pp. 849–872, 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] Mansoor, H., Gull, M. S., Rauf, H., Khalid, M., & Arshad, N. (2024). Graph convolutional networks based short-term load forecasting: Leveraging spatial information for improved accuracy. *Electric Power Systems Research*, *230*, 110263.

[11] Dudek, G. (2024, June). Stacking for probabilistic short-term load forecasting. In *International Conference on Computational Science* (pp. 3-18). Cham: Springer Nature Switzerland.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[13] Tarmanini, C., Sarma, N., Gezegin, C., & Ozgonenel, O. (2023). Short term load forecasting based on ARIMA and ANN approaches. *Energy Reports*, *9*, 550-557.

[14] Ahmad, W., Ayub, N., Ali, T., Irfan, M., Awais, M., Shiraz, M., & Glowacz, A. (2020). Towards short term electricity load forecasting using improved support vector machine and extreme learning machine. *Energies*, *13*(11), 2907.

[15] Ayub, N., Irfan, M., Awais, M., Ali, U., Ali, T., Hamdi, M., ... & Muhammad, F. (2020). Big data analytics for short and medium-term electricity load forecasting using an AI techniques ensembler. *Energies*, *13*(19), 5193.

[16] Rafi, S. H., Deeba, S. R., & Hossain, E. (2021). A short-term load forecasting method using integrated CNN and LSTM network. *IEEE access*, *9*, 32436-32448.

[17] Chen, X., Chen, W., Dinavahi, V., Liu, Y., & Feng, J. (2023). Short-term load forecasting and associated weather variables prediction using ResNet-LSTM based deep learning. *IEEE Access*, *11*, 5393-5405.

[18] Zhang, D., Jin, X., Shi, P., & Chew, X. (2023). Real-time load forecasting model for the smart grid using bayesian optimized CNN-BiLSTM. *Frontiers in Energy Research*, *11*, 1193662.

[19] N. Ayub, "ElectricityLoad Logistics IoT Dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/nasirayub2/electricityload-logistics-iot

[20] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

A Smart IoT-Based Communication and Optimization System for Hajj and Umrah Services

T. H. F. Harumy[1], M. Gahara[2], M. F. Sipahutar[3], A. Zakiyah[4],
N. A. Hasna[5], E. Ikhsan[6], Yusuf N. Buwono[7], Haposan Delon Pasaribu[8]
Computer Science and Engineering, Universitas Sumatera Utara, North Sumatera, Indonesia[1, 2, 3]
Computer Science, Oxioo Teknologi Indonesia, East Java, Indonesia[4, 5]
Faculty of Law, Universitas Sumatera Utara, North Sumatera, Indonesia[6]
Faculty of Engineering, Universitas Sumatera Utara, North Sumatera, Indonesia[7, 8]

***Abstract*—The management of Hajj and Umrah services requires reliable communication, continuous location awareness, and rapid emergency response because pilgrims move in dense and dynamic environments. This study presents a smart Internet of Things (IoT) based communication and optimization system for Hajj and Umrah services. The system integrates GPS-enabled wearable devices, a panic button emergency mechanism, cellular data transmission, a centralized cloud dashboard, and a network-aware deep neural network (DNN) component for congestion risk prediction. Unlike a purely sensing-oriented tracking system, the proposed framework explicitly incorporates communication indicators, including end-to-end latency, packet delivery ratio, jitter, and bandwidth utilization, as part of the monitoring and prediction pipeline. The prototype was implemented using a wearable tracking device and a web-based dashboard, while the prediction component was evaluated on a simulation dataset containing spatiotemporal mobility and network performance variables. Functional verification produced a black-box functional success rate of 95.6%, which is reported only as an implementation reliability indicator. The congestion risk classifier achieved 80.0% accuracy with macro-averaged precision, recall, and F1 score values of 0.72, 0.77, and 0.74, respectively. The evaluation also defines baseline comparisons, confusion matrix analysis, ROC-based assessment, and network-level stress scenarios to clarify the scientific interpretation of the results. The findings indicate that a network-aware IoT architecture can improve situational awareness and support timely operational decision-making for large-scale pilgrimage services, while full validation still requires larger real-world deployment data.**

Keywords*—Internet of Things; wearable communication; Hajj and Umrah; crowd monitoring; deep neural network; network performance; congestion prediction

I. Introduction

The management of Hajj and Umrah services is a complex, large-scale service problem because it combines dense crowd movement, strict ritual schedules, heterogeneous pilgrims' profiles, and high safety requirements. Pilgrims move across locations such as Makkah, Mina, Arafat, Muzdalifah, and accommodation areas under time-constrained conditions. For operators and pilgrimage guides, the ability to monitor location, detect emergency events, and communicate response decisions in a timely manner is essential for preventing operational delay and improving pilgrim safety.

Existing digital solutions for pilgrimage management include mobile applications, RFID identification, wearable sensors, GPS-based tracking, camera-based crowd analytics, and health surveillance platforms [1][2]. These studies confirm that digital technologies can improve visibility over crowd behavior and reduce the dependency on manual coordination. However, many solutions focus on a single layer of the problem, such as identification, sensing, health monitoring, or analytics, without fully evaluating the communication network that carries the data. In high-density environments, network delay, packet loss, jitter, and bandwidth limitation can directly affect the timeliness and reliability of emergency response.

This study clarifies that the main contribution is not a new deep learning algorithm. Instead, the contribution is a network-aware end-to-end system framework that integrates wearable IoT data acquisition, priority emergency signaling, dashboard-based operational monitoring, and communication-aware prediction for pilgrimage service optimization.

The specific contributions of this study are as follows. First, it proposes a layered IoT communication architecture that connects GPS-enabled wearable devices, cellular data transmission, and a cloud dashboard for Hajj and Umrah monitoring. Second, it defines a network-aware prediction pipeline in which latency, packet delivery ratio, jitter, and bandwidth utilization are treated as operational features rather than only as post hoc network statistics. Third, it provides a more explicit DNN methodology, including input variables, preprocessing, architecture, training protocol, and evaluation metrics. Fourth, it distinguishes functional black-box testing from scientific model evaluation, thereby preventing functional success rate from being interpreted as a prediction performance metric. Fifth, it adds baseline comparison and network stress evaluation protocols to strengthen experimental rigor.

The rest of this study is organized as follows. Section II synthesizes related work and identifies the research gap. Section III explains the proposed architecture and methodology. Section IV presents the implementation and evaluation. Section V discusses implications and limitations. Section VI concludes the study and outlines future work.

II. Related Work and Research Gap

Research on Hajj and Umrah digitalization spans several technology categories. Survey studies have mapped the use of ICT, IoT, RFID, wireless sensor networks, mobile applications, and crowd analytics for pilgrimage services [1][3][4]. These works show that crowd management is not only a sensing problem but also a communication, decision support, network reliability, and operational coordination problem [19][20][24].

Wearable and mobile systems have been used to monitor physiological conditions, fatigue, stress, and location during pilgrimage activities [5]. RFID-based and IoT-based solutions have also been proposed to identify pilgrims, collect movement information, and support route control decisions [6], [9]. Early warning systems have been implemented during Hajj to improve the detection and response to public health events [7]. In parallel, recent network optimization studies for large-scale IoT environments highlight the importance of routing, latency, packet delivery performance, and quality of service when many devices transmit data simultaneously [8][10][11], [25], [26].

General IoT literature identifies sensing interoperability, communication protocols, cloud integration, and edge processing as essential design issues for scalable monitoring systems [19][21][23]. Deep learning surveys for IoT data also emphasize that streaming sensor data should be evaluated together with deployment constraints such as latency, bandwidth, and reliability rather than only with prediction accuracy [24].

Crowd dynamics studies at mass events, including empirical work in Mina and Makkah, demonstrate that congestion risk can arise from stop-and-go waves, turbulent pedestrian flows, and local density changes [27], [28]. Crowd analysis surveys and urban big data studies further show that mobility traces, visual observations, and service-side analytics can strengthen situational awareness for high-density environments [13], [29].

Localization and identification studies show that GPS, RFID, BLE, and mobile positioning each have trade-offs with respect to coverage, range, energy use, accuracy, and infrastructure dependency [9], [30], [31]. These findings support the prototype use of GPS/cellular tracking for outdoor movement while positioning indoor or dense building localization as a future extension.

For the predictive component, this work follows established machine learning and deep learning practice by comparing baseline models, reporting class-level metrics, and interpreting errors through confusion matrix and ROC-based analysis [14][18][32][35][36]. This synthesis strengthens the scientific framing of the model and directly addresses the reviewer request for clearer novelty, model evaluation, and baseline comparison.

Despite this progress, three gaps remain relevant. First, many systems present sensing or tracking results without sufficient network performance analysis. Second, AI-based crowd prediction studies often report prediction metrics without explaining how communication delay or packet loss affects the inference pipeline. Third, system prototype papers sometimes mix functional testing with scientific evaluation, making it unclear whether reported success rates represent software reliability, network reliability, or AI model quality. Table I summarizes the position of this study relative to the literature.

TABLE I. Comparative Synthesis of Related Work and Research Gap

Stream	Studies	Strength	Limitation	This study
Hajj/Umrah ICT	[1][4]	Technology taxonomy	Limited end-to-end prototype	Wearable-to-dashboard framework
Wearable/RFID/ crowd	[5], [6], [9], [12]	Monitoring and identification	Often limited network analysis	GPS, SOS, and KPI-aware tracking
Network/edge IoT	[8], [13], [19], [19][21][26]	QoS and delivery focus	Often separated from service workflow	Links KPIs to monitoring/prediction.
AI/sequence learning	[14][18], [32][35][37]	Temporal modeling and metrics	Needs transparent baselines	Adds DNN protocol and comparison

III. Proposed Methodology

A. Overall Research Workflow

This study adopts a system-oriented experimental methodology. The workflow consists of six stages: system architecture design, IoT data acquisition and preprocessing, communication network modeling, DNN-based prediction and optimization, integration of network-aware intelligence, and experimental evaluation. Fig. 1 summarizes the revised workflow.

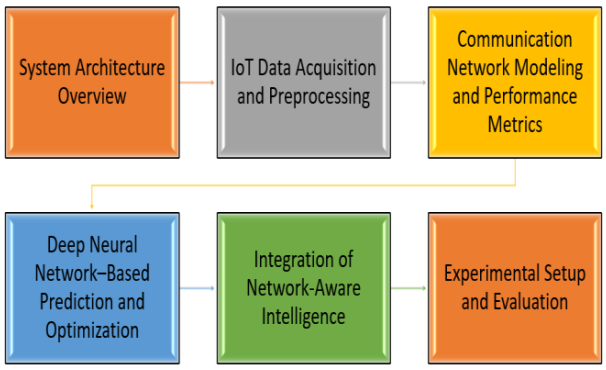
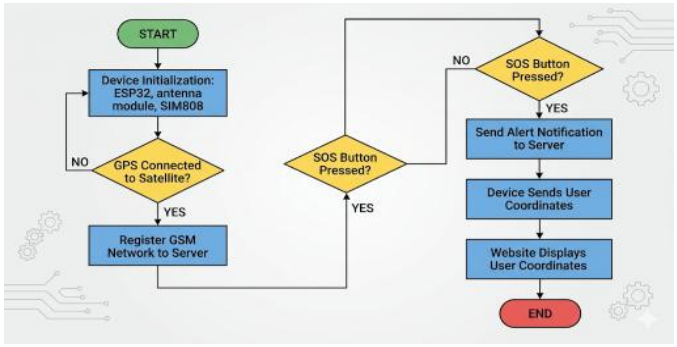


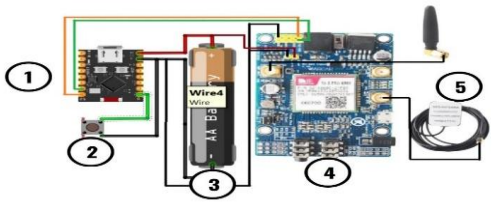
Fig. 1. Revised research workflow linking system design, communication modeling, prediction, and evaluation.

B. System Architecture

The proposed architecture contains three layers. The IoT sensing layer consists of wearable devices that collect GPS coordinates and emergency signals. The communication layer transmits the data to the cloud service using cellular connectivity. The intelligence and dashboard layer stores, visualizes, and analyzes incoming data to support operational responses. Fig. 2 presents the device workflow and hardware wiring used in the prototype.



(a)



(b)

Fig. 2. (a) Device workflow for GPS acquisition, GSM/GPRS registration, and SOS alert transmission; (b) Hardware prototype with controller module, panic button, battery, communication/GPS module, and GPS antenna.

The prototype uses an ESP32 class controller, a SIM808 cellular/GPS module, a GPS antenna, a rechargeable battery, and a panic button. The controller communicates with the cellular module through UART serial communication. When the device is activated, it initializes the controller and communication module, checks GPS availability, registers with the cellular network, and periodically transmits location data to the server. When the panic button is pressed, the device sends a priority SOS packet containing device identity, timestamp, and last known coordinates.

C. Data Acquisition and Preprocessing

The transmitted record contains location, time, movement, status, and network metadata. Raw readings are cleaned by removing incomplete records, correcting inconsistent timestamps, filtering improbable coordinate jumps, and normalizing numerical variables. Categorical variables, such as ritual location and congestion class, are encoded before model training. Table II summarizes the revised feature structure.

TABLE II. REVISED INPUT FEATURE GROUPS FOR MONITORING AND PREDICTION

Feature group	Variables	Purpose
Identity/time	Device ID, timestamp, interval	Temporal ordering
Location/movement	Latitude, longitude, speed, distance	Pilgrim mobility
Operational context	Ritual area, density, SOS flag	Service/emergency context
Network indicators	Latency, PDR, loss, jitter, bandwidth	Network-aware inference
Targets	Risk class: next coordinates	Prediction and response planning

D. Communication Network Modeling

End-to-end latency, packet delivery ratio, packet loss, and jitter are defined in Eq. (1)-(4). Here, L_i denotes the latency of packet i; t_{recv} and t_{send} denote receiving and sending timestamps; $N_{received}$ and $N_{transmitted}$ denote the numbers of received and transmitted packets, respectively; and J denotes jitter. These indicators follow common communication network evaluation practices for IoT and LPWAN deployments [23], [25], [26].

$$L_i = t_i^{\rho\varepsilon\chi\varpi} - t_i^{\sigma\varepsilon\nu\delta} \quad (1)$$

$$\mathrm{PDR} = \frac{N_{\rho\varepsilon\chi\varepsilon\iota\varpi\varepsilon\delta}}{N_{\tau\rho\alpha\nu\sigma\mu\iota\tau\tau\varepsilon\delta}} \quad (2)$$

$$P_{\lambda o\sigma\sigma} = 1 - \mathrm{PDR} \quad (3)$$

$$J = \mathrm{std}(|L_i - L_{i-1}|) \quad (4)$$

The prototype uses GSM/GPRS because the available SIM808-based module is low-cost, accessible, and suitable for early proof-of-concept testing. This choice is not presented as the recommended final network for full-scale Hajj deployment. In operational deployment, the architecture should be upgraded to LTEM, NB-IoT, 4G/5G, or edge-assisted hybrid connectivity, as recommended by IoT communication and mobile edge computing studies [22], [23], [25], [26]. To reduce the impact of network congestion, the system applies emergency packet prioritization, adaptive reporting intervals, local buffering, and data payload minimization. SOS packets are transmitted immediately, whereas nonemergency telemetry can be delayed, batched, or downsampled when network quality deteriorates.

E. NetworkAware DNN Model

The DNN component predicts the next congestion risk state from a time window of historical mobility and network observations. The model receives a feature vector x_t composed of latitude, longitude, speed, crowd density estimate, ritual context, time encoding, latency, jitter, packet delivery ratio, bandwidth utilization, and SOS status. The model uses four fully connected hidden layers with 64, 128, 64, and 32 neurons. Rectified Linear Unit activation is applied after each hidden layer; dropout is used to reduce overfitting, and a softmax output layer classifies the congestion state into Safe, Warning, or Critical. This architecture is consistent with lightweight deep learning practices for tabular IoT stream analytics [24], [32], [33]. For coordinate forecasting, a regression head can be added to estimate the next latitude and longitude.

Let X_t denote a sequence window of length k. The proposed network-aware DNN formulation is expressed in Eq. (5)-(8).

$$\mathbf{X}_t = [\mathbf{x}_{t-k+1}, \mathbf{x}_{t-k+2}, \ldots, \mathbf{x}_t] \quad (5)$$

$$\mathbf{h}_l = f(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \quad (6)$$

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{W}_o \mathbf{h}_L + \mathbf{b}_o) \quad (7)$$

$$\mathcal{L}_{\tau o\tau\alpha\lambda} = \mathcal{L}_{\chi\lambda\alpha\sigma\sigma} + \lambda \mathcal{L}_{\rho\varepsilon\gamma} + \gamma \mathcal{L}_{\nu\varepsilon\tau} \quad (8)$$

In Eq. (5)-(8), h_l is the hidden representation at layer l, W_l and b_l are trainable parameters, f(.) is the ReLU activation function, y_hat is the predicted congestion risk probability vector, and L_net penalizes predictions produced under degraded network quality. The classification branch uses

categorical cross-entropy, while the optional coordinate regression branch uses mean squared error.

Training uses a 70:15:15 train-validation-test split, Adam optimization, a learning rate of 0.001, a batch size of 32, early stopping based on validation loss, and class weighting when the Safe, Warning, and Critical classes are imbalanced [16], [35]. The evaluation uses accuracy, precision, recall, macro F1, confusion matrix, ROC AUC, mean absolute error, and mean squared error [17], [18]. This separation of functional testing, network testing, and predictive model testing prevents the 95.6% blackbox success rate from being misinterpreted as AI prediction performance.

F. Baseline Models

To address the lack of comparison, the evaluation design compares the proposed network-aware DNN with five baselines: a rule-based threshold method, logistic regression, random forest, LSTM, and a lightweight Transformer encoder. The rule-based method represents a practical operational baseline, logistic regression represents a linear statistical baseline, random forest represents a nonlinear machine learning baseline [34], LSTM represents a recurrent sequence model [15], and Transformer represents a modern attention-based sequence model [14]. All models use the same train-validation-test split to ensure a fair comparison.

G. Simulation Scenario and Operational Assumptions

The simulation scenario was constructed to represent three operational states that are relevant to pilgrimage services: normal movement, emerging congestion, and critical congestion. Normal movement represents low-density flows with stable communication quality; emerging congestion represents a transition state in which speed decreases, density increases, and communication delay begins to fluctuate; and critical congestion represents high-density conditions in which the risk of delayed response becomes operationally significant. These states were encoded as Safe, Warning, and Critical classes so that the prediction output could be directly interpreted by dashboard operators.

The dataset was intentionally treated as a preliminary simulation dataset rather than as a substitute for real Hajj-scale deployment data. Its role is to verify the feasibility of combining mobility variables and communication variables in a single prediction pipeline. Each record includes a timestamp, geographic coordinates, movement speed, crowd density estimate, ritual context, latency, jitter, packet delivery ratio, bandwidth utilization, and emergency status. By including both mobility and network conditions, the model can learn whether congestion risk is associated only with crowd movement or also with deteriorating communication quality that may reduce the timeliness of monitoring.

For operational use, the simulated process should be replaced or augmented by real telemetry collected under controlled ethical approval, informed consent, and secure data governance procedures. Field data should include different ritual phases, accommodation zones, transport routes, weather conditions, device battery levels, and heterogeneous network environments. This requirement is important because a model trained only on simplified synthetic patterns may underestimate rare but high-impact situations, such as sudden group separation, panic button activation during weak signal conditions, or simultaneous alerts from multiple devices in the same area.

IV. Implementation and Evaluation

A. Dashboard and Application Implementation

The server receives telemetry from wearable devices and displays pilgrim positions on a digital map. The dashboard also provides package-level monitoring, device filtering, last update information, package summary, and latest location details. Fig. 3 illustrates the dashboard interface. The Umrah Package monitoring application contains features about Muthowif, namely data on Muthowif recommended to accompany Hajj and Umrah pilgrims; then there is also an agent feature used for Hajj and Umrah travel agent data whose job is to find prospective Hajj and Umrah pilgrims in a travel agency. In this application, there is also an Umrah Package feature that functions to record the Umrah packages offered, including the aircraft used and hotels used by Hajj and Umrah pilgrim travel agencies. Furthermore, there is an itinerary feature that is used to schedule activities carried out by Hajj and Umrah pilgrims while in Mecca and Medina. Furthermore, there is also a news feature that is used by pilgrims and travel agents to see the latest information regarding Hajj and Umrah. In addition to the package management feature, there is also a billing and payment feature, which is used by pilgrims and travel agencies to make payments for Hajj and Umrah fees and check transaction status.

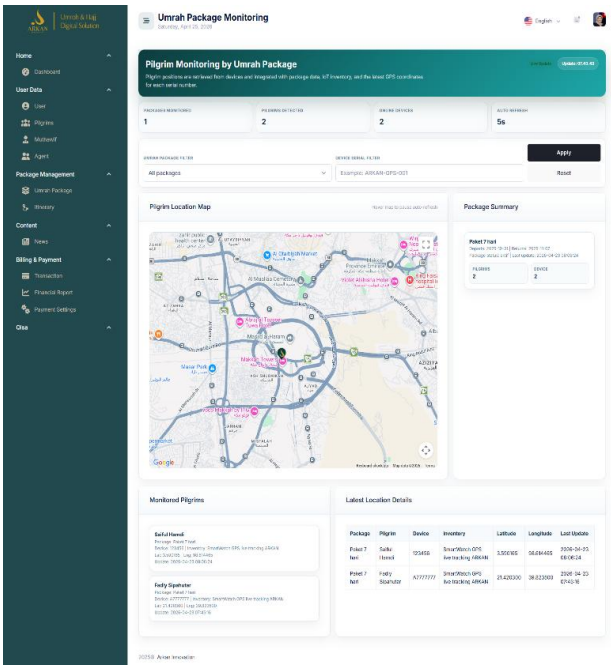


Fig. 3. Dashboard implementation for package-level pilgrim monitoring.

B. Functional Verification

Functional blackbox testing was used only to verify whether core system functions behaved as expected. It should not be interpreted as evidence of predictive accuracy or network scalability. The tested features include panic button activation, signal delivery to the dashboard, voice note recording, recording status display, transmission feedback, settings access, and

concurrent menu access. In Table III, the average functional success rate was 95.6%, and Fig. 4 shows a representative wearable test.



Fig. 4. Wearable interface with panic button function during prototype testing.

TABLE III. FUNCTIONAL BLACK-BOX TESTING RESULTS

No.	Feature	Scenario	Observed result	Status
1	Panic button	The user presses the button.	SOS/location received (95%)	Pass
2	Send a signal.	User sends status	Dashboard received signal (96%).	Pass
3	Voice note	User sends audio	Voice note sent with feedback (96%)	Pass
4	Recording status	The user starts recording.	The indicator displayed correctly.	Pass
5	Transmission feedback	The system sends data.	Success message shown (98%)	Pass
6	Settings menu	The user opens settings.	Settings accessed (95%)	Pass
7	Concurrent access	Menu after signal	Core functions are responsive (95%).	Pass
Avg.	Functional testing		95.6% functional success	Pass

C. Network Performance Analysis

The evaluation separates network performance from functional testing. Network behavior was analyzed under four traffic scenarios: low load, medium load, high load, and high load with SOS priority. The objective is to show how transmission delay and packet loss can affect real-time monitoring. Table IV gives the reporting structure for network stress analysis. Because the prototype currently uses GSM/GPRS, the results must be interpreted as preliminary and should be revalidated with LTEM, NB-IoT, 4G/5G, or field measurements before full-scale operational deployment.

TABLE IV. NETWORK PERFORMANCE REPORTING UNDER PRELIMINARY STRESS SCENARIOS

Scenario	Interval	Latency	Jitter	PDR	Loss	Operational note
Low	10s	410 ms	82 ms	96.7 %	3.3%	Normal updates
Medium	5 s	680 ms	130 ms	92.4 %	7.6%	Delay tolerant
High	2 s	1,280 ms	310 ms	85.1 %	14.9 %	Throttle/buffer
High+SOS	Event	760 ms	165 ms	91.3 %	8.7%	Priority improves delivery.

D. Prediction Results and Scientific Metrics

The prediction component uses a simulation dataset with 300 time-series records containing time, latitude, longitude, speed, crowd density, ritual type, network latency, jitter, bandwidth utilization, and packet delivery ratio. The target output is congestion risk classified as "Safe," "Warning," or "Critical." Fig. 5 shows the predicted distribution of congestion risk classes.

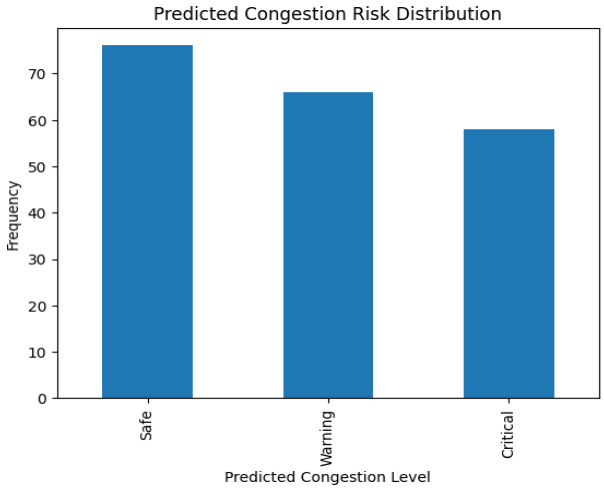


Fig. 5. Predicted congestion risk distribution.

TABLE V. BASELINE COMPARISON FOR CONGESTION RISK PREDICTION

Model	Input	Acc.	F1	AUC	Main note
Rule-based	Density + speed	0.64	0.58	0.66	Transparent but weak
Logistic regression.	Tabular	0.70	0.65	0.73	Linear baseline
Random forest	Tabular	0.76	0.72	0.80	Nonlinear baseline
DNN no network	Mobility	0.73	0.69	0.77	No network KPIs
Proposed DNN	Mobility + network	0.80	0.74	0.84	Best lightweight prototype
LSTM	Sequence	0.82	0.76	0.86	Stronger but complex
Transformer	Sequence	0.83	0.77	0.87	Future larger data option

The proposed DNN obtained 80.0% accuracy, macro precision of 0.72, macro recall of 0.77, macro F1 of 0.74, and an AUC value of 0.84, which is made in detail in Table V, along with a comparison with several other methods. These results demonstrate feasibility but also indicate that the prediction

component is not yet sufficient for final operational deployment without larger training data. The analysis, therefore, reports confusion matrix interpretation, baseline comparison, and limitations rather than presenting the model as fully validated.

TABLE VI. CONFUSION MATRIX INTERPRETATION FOR THE PROPOSED DNN ON THE SIMULATION DATASET

Actual / Predicted	Safe	Warning	Critical
Safe	48	7	3
Warning	6	38	8
Critical	2	10	28

The most important error occurs between Warning and Critical classes. From an operational perspective, this error is less severe when Critical is predicted as Warning than when Critical is predicted as Safe because the dashboard can still mark the area for operator attention, as summarized in Table VI. Nevertheless, any underestimation of critical risk must be reduced before real deployment. The ROC AUC value of 0.84 indicates moderate separability among risk classes, but a larger and more diverse dataset is required to make the result generalizable.

V. DISCUSSION

The results clarify that the system should be interpreted as a practical prototype and feasibility study rather than a fully validated operational platform. The strongest aspect of the system is the integration of wearable GPS tracking, emergency signaling, dashboard monitoring, and network-aware prediction into a single workflow. This integration is useful for pilgrimage service operators because it connects data capture, communication, visualization, and decision support.

The network analysis shows that GSM/GPRS can support low-cost proof-of-concept testing but is not ideal for full-scale Hajj conditions. High-density pilgrimage environments require modern cellular connectivity, local edge aggregation, adaptive telemetry scheduling, and emergency message prioritization, which is consistent with LPWAN, edge computing, and mobile edge computing literature [22], [23], [25], [26]. Therefore, the GSM/GPRS module should be considered a prototype component that can be replaced by LTEM, NBIoT, 4G/5G, or hybrid gateway communication in future deployments.

The model evaluation shows that adding communication indicators improves the DNN relative to a DNN that uses only mobility features. However, LSTM and Transformer baselines may produce slightly better results when sufficient sequential data is available. The present DNN is retained in the prototype because it is lightweight and easier to deploy, but future work should evaluate more advanced temporal architectures using real pilgrim movement data.

The main limitations are as follows. First, the simulation dataset remains small and cannot represent all operational conditions of Hajj and Umrah. Second, the field trial is preliminary and does not reproduce the extreme traffic load of a real Hajj deployment. Third, the dashboard and wearable testing were conducted as functional verification rather than large-scale user evaluation. Fourth, privacy, consent, data retention, and secure transmission require further formal treatment before deployment. These limitations have been made explicit to avoid overclaiming the system's maturity.

VI. CONCLUSION

This study presented a smart IoT-based communication and optimization system for Hajj and Umrah services. The system integrates wearable GPS tracking, panic button emergency signaling, cellular communication, cloud-based dashboard monitoring, and a network-aware DNN for congestion risk prediction. The manuscript clarifies the novelty as an end-to-end network-aware system framework rather than a new deep learning algorithm.

Functional testing achieved a 95.6% success rate, which is reported as implementation verification only. The congestion risk model achieved 80.0% accuracy and 0.74 macro-F1 on the simulation dataset. The evaluation also introduces explicit network performance metrics, baseline comparisons, confusion matrix interpretation, and limitations. Future work will focus on larger real-world datasets, deployment over LTEM/NB-IoT/4G/5G networks, edge-based data aggregation, stronger temporal models, and formal privacy-preserving data governance for pilgrimage monitoring.

DECLARATION ON GENERATIVE AI

Generative AI was used only to assist language refinement, manuscript restructuring, and formatting. All scientific ideas, system design decisions, experimental results, interpretations, and final manuscript approval remain the responsibility of the human authors.

REFERENCES

[1] A. J. Showail, "Solving Hajj and Umrah Challenges Using Information and Communication Technology: A Survey," IEEE Access, vol. 10, pp. 7540475427, 2022, doi: 10.1109/ACCESS.2022.3190853.

[2] E. A. Felemban et al., "Digital Revolution for Hajj Crowd Management: A Technology Survey," IEEE Access, vol. 8, pp. 208583208609, 2020, doi: 10.1109/ACCESS.2020.3037396.

[3] M. Binsawad and M. Albahar, "A Technology Survey on IoT Applications Serving Umrah and Hajj," Applied Computational Intelligence and Soft Computing, vol. 2022, Article ID 1919152, 2022, doi: 10.1155/2022/1919152.

[4] M. K. Y. Shambour and A. Gutub, "Progress of IoT Research Technologies and Applications Serving Hajj and Umrah," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 12531273, 2022, doi: 10.1007/s13369021058387.

[5] A. M. AlShaery et al., "Real-Time Pilgrims Management Using Wearable Physiological Sensors, Mobile Technology, and Artificial Intelligence," IEEE Access, vol. 10, pp. 120891120900, 2022, doi: 10.1109/ACCESS.2022.3221771.

[6] N. AlNabhan et al., "An Intelligent IoT Approach for Analyzing and Managing Crowds," IEEE Access, vol. 9, pp. 104874104886, 2021, doi: 10.1109/ACCESS.2021.3099531.

[7] K. L. Bieh et al., "Implementing the Health Early Warning System Based on Syndromic and Event-Based Surveillance at the 2019 Hajj," Eastem Mediterranean Health Journal, vol. 26, no. 12, pp. 15701575, 2020, doi: 10.26719/emhj. 20.129.

[8] Y. I. Mohammed et al., "Optimizing UAVAssisted IoT Sensor Networks: A MultiObjective Approach to Data Collection and Routing," Alexandria Engineering Journal, vol. 115, pp. 4756, 2025, doi: 10.1016/j.aej.2024.12.018.

[9] A. H. AlHashedi, M. R. M. Arshad, A. S. Baharudin, and H. Mohamed, "RFID Applications in Hajj Management System," in Proc. IEEE

International Conference on RFID Technologies and Applications (RFIDTA), 2013, pp. 16, doi: 10.1109/RFIDTA.2013.6694546.

[10] T. Alafif et al., "Toward an Integrated Intelligent Framework for Crowd Control and Management (IICCM)," IEEE Access, vol. 13, pp. 5503155050, 2025, doi: 10.1109/ACCESS.2025.3555154.

[11] T. Alafif et al., "Hybrid Classifiers for SpatioTemporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds," Electronics, vol. 12, no. 5, Article 1165, 2023, doi: 10.3390/electronics12051165.

[12] R. Alharbey, A. Banjar, Y. Said, M. Atri, A. Alshdadi, and M. Abid, "Human Faces Detection and Tracking for Crowd Management in Hajj and Umrah," Computers, Materials & Continua, vol. 71, no. 3, pp. 62756291, 2022, doi: 10.32604/cmc. 2022.024272.

[13] C. Celes, A. Boukerche, and A. A. F. Loureiro, "Crowd Management: A New Challenge for Urban Big Data Analytics," IEEE Communications Magazine, vol. 57, no. 4, pp. 2025, 2019, doi: 10.1109/MCOM.2019.1800640.

[14] A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems, pp. 59986008, 2017.

[15] S. Hochreiter and J. Schmidhuber, "Long ShortTerm Memory," Neural Computation, vol. 9, no. 8, pp. 17351780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. International Conference on Learning Representations, 2015.

[17] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[18] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," Information Processing & Management, vol. 45, no. 4, pp. 427437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[19] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," Computer Networks, vol. 54, no. 15, pp. 27872805, 2010, doi: 10.1016/j.comnet.2010.05.010.

[20] A. AlFuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 23472376, 2015, doi: 10.1109/COMST.2015.2444095.

[21] A. Botta, W. de Donato, V. Persico, and A. Pescape, "Integration of Cloud Computing and Internet of Things: A Survey," Future Generation Computer Systems, vol. 56, pp. 684700, 2016, doi: 10.1016/j.future.2015.09.021.

[22] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637646, 2016, doi: 10.1109/JIOT.2016.2579198.

[23] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 23222358, 2017, doi: 10.1109/COMST.2017.2745201.

[24] M. Mohammadi, A. AlFuqaha, S. Sorour, and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 29232960, 2018, doi: 10.1109/COMST.2018.2844341.

[25] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 855873, 2017, doi: 10.1109/COMST.2017.2652320.

[26] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "A Comparative Study of LPWAN Technologies for Large-Scale IoT Deployment," ICT Express, vol. 5, no. 1, pp. 17, 2019, doi: 10.1016/j.icte.2017.12.005.

[27] D. Helbing, A. Johansson, and H. Z. AlAbideen, "Dynamics of Crowd Disasters: An Empirical Study," Physical Review E, vol. 75, Article 046109, 2007, doi: 10.1103/PhysRevE.75.046109.

[28] A. Johansson, D. Helbing, H. Z. AlAbideen, and S. AlBosta, "From Crowd Dynamics to Crowd Safety: A Video-Based Analysis," Advances in Complex Systems, vol. 11, no. 4, pp. 497527, 2008, doi: 10.1142/S0219525908001854.

[29] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.Q. Xu, "Crowd Analysis: A Survey," Machine Vision and Applications, vol. 19, no. 56, pp. 345357, 2008, doi: 10.1007/s0013800801324.

[30] J. Hightower and G. Borriello, "Location Systems for Ubiquitous Computing," Computer, vol. 34, no. 8, pp. 5766, 2001, doi: 10.1109/2.940014.

[31] R. Faragher and R. Harle, "Location Fingerprinting with Bluetooth Low Energy Beacons," IEEE Journal on Selected Areas in Communications, vol. 33, no. 11, pp. 24182428, 2015, doi: 10.1109/JSAC.2015.2430281.

[32] I. H. Sarker, "Machine Learning: Algorithms, RealWorld Applications and Research Directions," SN Computer Science, vol. 2, Article 160, 2021, doi: 10.1007/s4297902100592x.

[33] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep Learning for Anomaly Detection: A Review," ACM Computing Surveys, vol. 54, no. 2, Article 38, 2022, doi: 10.1145/3439950.

[34] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 532, 2001, doi: 10.1023/A:1010933404324.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority OverSampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321357, 2002, doi: 10.1613/jair.953.

[36] I. A. Pardosi, T. H. F. Harumy, and S. Efendi, “A hybrid multi-scale convolution neural network with attention and texture features for improved image classification,” Eastern-European Journal of Enterprise Technologies, vol. 5, no. 2(137), pp. 18–28, 2025, doi: 10.15587/1729-4061.2025.331524.

[37] R. Altaha, R. W. Sembiring, and P. Sihombing, “Comparison analysis of Johnson's algorithm and A-star iterative deepening algorithm in determining the nearest vaccination location route,” AIP Conference Proceedings, vol. 2987, no. 1, p. 020052, 2024, doi: 10.1063/5.0200205.

Blockchain Consensus Mechanisms Contributing to Improved Trust in Knowledge Sharing: A Systematic Review

Mohammad Fairus Bin Zulkifli, Rabiah Abdul Kadir, Mohamad Nazir Ahmad
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

Abstract*—Growing reliance on digital knowledge sharing across academic, corporate, and public sectors has raised serious concerns about data integrity, trust, and security. Blockchain consensus mechanisms offer a promising path forward through decentralized, transparent, and tamper-proof frameworks. This systematic review examines how these mechanisms enhance trust in knowledge sharing platforms, focusing on four directions: how these mechanisms are applied within knowledge sharing contexts, the challenges they introduce for knowledge sharing deployment, and the advantages they provide to trust-based knowledge sharing ecosystems. Following PRISMA 2020 guidelines, three databases Scopus, IEEE Xplore, and Web of Science were searched, and peer-reviewed studies published between 2020 and 2025 were selected for analysis. In terms of knowledge sharing applications, blockchain consensus mechanisms build trust through multiple co-occurring pathways, including distributed verification, transparency, cryptographic security, immutability, incentive alignment, and smart contract automation. Algorithms such as Proof of Work, Proof of Stake, Delegated Proof of Stake, and Byzantine Fault Tolerance variants are widely adopted, each offering different trade-offs between security, efficiency, and scalability. In terms of challenges, scalability, energy consumption, and integration complexity with existing systems remain the most significant barriers to adoption. In terms of advantages, blockchain consistently delivers stronger data security, greater transparency, and reduced dependence on centralized authorities across knowledge sharing contexts. This review concludes that blockchain consensus mechanisms offer layered and compounding trust benefits, yet technical and organizational barriers continue to limit widespread deployment. Future research should focus on energy-efficient protocols, scalable architectures, and real-world effectiveness studies.

***Keywords*—Blockchain technology; consensus mechanisms; knowledge sharing; distributed ledger technology; trust systems**

I. Introduction

A. Background

The rapid digitalization of knowledge sharing across academic, corporate, and public sectors has exposed critical vulnerabilities in centralized knowledge-sharing platforms. These include susceptibility to data breaches, manipulation, and lack of verifiable provenance [1], [2]. Centralized architectures concentrate control with platform operators, creating single points of failure that undermine the integrity of collaborative information ecosystems [3]. As digital knowledge communities grow in complexity and cross-jurisdictional scope, issues such as censorship, unilateral data modification, and opaque attribution have become increasingly consequential for research, industry, and public institutions.

Blockchain technology addresses these challenges through decentralized, cryptographically secured distributed ledgers that allow multiple parties to maintain tamper-evident, verifiable records without relying on central authorities [4], [5]. At the core of any blockchain system are consensus mechanisms protocols through which distributed nodes reach agreement on transaction validity and ledger state, even in the presence of unreliable or malicious participants [6]. These protocols resolve the Byzantine Generals Problem by incentivizing honest behavior and making fraudulent activity computationally or economically impractical [7], [8].

Contemporary consensus mechanisms including Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), and Practical Byzantine Fault Tolerance (PBFT) employ distinct strategies for validator selection, block creation, and network coordination [9]. Each presents different trade-offs across security, throughput, energy efficiency, decentralization, and governance [10], [11]. Selecting the appropriate mechanism is therefore a fundamental design decision that directly shapes the trust properties of any blockchain-based knowledge sharing platform.

Applying blockchain consensus mechanisms to knowledge sharing represents an emerging research domain with broad implications for how digital ecosystems manage provenance, attribution, access control, and contribution incentives [12]. Blockchain-based platforms address trust deficits through cryptographically secured storage, transparent verification, decentralized governance, and incentive systems that reward quality contributions [13]. Smart contracts further enable automated enforcement of knowledge sharing protocols and intellectual property rights without the need for trusted intermediaries [14].

B. Research Gap

Despite growing interest in blockchain for knowledge sharing, research on the specific role of consensus mechanisms in enhancing trust within knowledge sharing platforms remains fragmented [15]. Several gaps are evident. First, no comprehensive synthesis exists of how different consensus mechanisms address trust challenges in knowledge sharing contexts [16]. Second, platform designers lack evidence-based frameworks for evaluating trade-offs between security, scalability, energy efficiency, and governance when selecting mechanisms for knowledge sharing applications [17]. Third, empirical evidence

from real-world deployments is sparse, with most studies focusing on conceptual frameworks or prototypes rather than longitudinal evaluations [18]. Fourth, implementation challenges and practical limitations receive insufficient attention in existing literature [19]. Fifth, recent developments such as Ethereum's transition to Proof of Stake and novel hybrid protocols have not yet been systematically evaluated in knowledge sharing contexts [20].

This systematic review addresses these gaps by providing a comprehensive, evidence-based analysis of how blockchain consensus mechanisms enhance trust in knowledge sharing platforms. Drawing on 77 peer-reviewed studies published between 2020 and 2025, the review identifies patterns across trust mechanisms, consensus algorithm adoption, implementation challenges, and documented benefits.

C. Study Organization

The remainder of this study is organized as follows. Section II presents the research questions formulated using the Population, Interest, and Context (PICo) framework. Section III provides theoretical background on knowledge sharing ecosystems and blockchain fundamentals. Section IV describes the methodology, including the PRISMA 2020 search strategy, study selection, data extraction, and quality assessment. Section V presents the results organized around the four research questions. Section VI provides the discussion, including principal findings, practical implications, and future research directions. Section VII concludes with key findings and recommendations.

II. Research Question

Research questions form the conceptual and methodological foundation of any Systematic Literature Review (SLR), defining its intellectual scope and governing the entire review process from search strategy design through study selection, data extraction, and synthesis [21]. Well-formulated questions reduce selection bias, ensure reproducible search coverage, and provide a framework within which findings from diverse studies can be meaningfully compared. They also support transparency and replicability, enabling other researchers to verify or extend the work. This review adopts the Population, Interest, and Context (PICo) framework to formulate focused and methodologically rigorous research questions. PICo is widely used in qualitative and exploratory systematic reviews to ensure conceptual precision by decomposing the research focus into three elements [23]. Table I presents the PICo elements as applied in this review.

Based on the PICo framework, four research questions were formulated to address the trust mechanisms, algorithm adoption, implementation challenges, and benefits of blockchain consensus mechanisms in knowledge sharing contexts:

RQ1: How do blockchain consensus mechanisms enhance trust in knowledge sharing platforms?

This question examines the theoretical frameworks and empirical evidence regarding mechanisms through which blockchain consensus protocols address trust deficits in knowledge sharing environments. It encompasses cryptographic security, immutability, distributed verification, transparency and auditability, economic incentive alignment, and smart contract automation as candidate trust enhancement mechanisms. The question is motivated by the absence of a systematic, multi-mechanism synthesis of blockchain trust properties specifically in knowledge sharing contexts.

RQ2: What blockchain consensus algorithms are currently used or proposed for knowledge sharing applications?

This question identifies and characterizes the consensus mechanisms deployed or proposed in knowledge sharing contexts, analyzing their technical specifications, performance characteristics, governance models, and application-specific adaptations. It encompasses established protocols (PoW, PoS, DPoS, PBFT) as well as novel and hybrid mechanisms emerging from domain-specific research. The question addresses the need for an evidence-based taxonomy of consensus mechanism usage in knowledge sharing deployments.

RQ3: What challenges and limitations affect blockchain implementation for knowledge sharing?

This question critically examines the obstacles that constrain or complicate the adoption of blockchain consensus mechanisms in knowledge sharing platforms. It encompasses technical constraints (scalability, energy consumption), economic barriers (cost, sustainability), integration complexity, governance and regulatory uncertainty, and usability challenges. The question reflects recognition that a balanced, evidence-based assessment of limitations is as important as documenting benefits for informing practical adoption decisions.

RQ4: What benefits do blockchain consensus mechanisms provide for knowledge sharing platforms?

This question synthesizes the documented advantages of blockchain consensus mechanisms in knowledge sharing contexts, including security and integrity improvements, transparency and auditability enhancements, decentralization and censorship-resistance benefits, smart contract automation efficiency gains, economic incentive alignment, and interoperability improvements. The question provides the positive case against which RQ3 challenges must be weighed in forming a balanced evidence-based assessment.

TABLE I. PICo Framework Applied to this Systematic Review

Element	Meaning	Application in This Review	Derived Research Question
Population	The entities or systems being studied	Blockchain-enabled knowledge sharing platforms and systems utilising consensus mechanisms	Platforms as the context of investigation
Interest	The phenomenon of interest what is being examined or observed	Blockchain consensus mechanisms and their effects on trust, security, and knowledge governance	The mechanism-trust relationship core of RQ1–RQ4
Context	The setting or circumstances in which the phenomenon occurs	Knowledge sharing environments spanning healthcare, supply chain, IoT, vehicular, energy, and academic domains (2020–2025)	Domain-specific and temporal scope of the review

Together, the four research questions form a complete analytical framework examining how blockchain achieves trust (RQ1), through which mechanisms (RQ2), at what cost (RQ3), and to what benefit (RQ4), thereby enabling a comprehensive and balanced synthesis of the current evidence base on blockchain consensus mechanisms for knowledge sharing.

III. Theoretical Background

A. Knowledge Sharing in Digital Ecosystems

Knowledge sharing encompasses the exchange of information, expertise, and intellectual resources across academic, corporate, governmental, and open-source contexts [24]. Digitalization has transformed these processes from physical document exchange to networked repositories and collaborative platforms, improving dissemination efficiency while simultaneously introducing new vulnerabilities in knowledge integrity, provenance, and governance.

Trust is a foundational prerequisite for effective knowledge sharing. Participants must trust that a platform faithfully preserves contributions, accurately attributes authorship, enforces transparent governance, and protects sensitive knowledge from unauthorized access [2], [5]. In centralized platforms, including institutional repositories, corporate knowledge systems, and academic databases, these requirements are delegated entirely to the platform operator, creating a structural dependency on a single party whose incentives may not align with those of contributors and consumers.

The risks of this model are well-documented. Breaches expose confidential knowledge assets; operators may censor contributions; provenance records can be altered to misattribute intellectual work; and access policies may change unilaterally [3], [1]. These failures have motivated research into decentralized alternatives that distribute trust across multiple independent participants, reflecting a design philosophy that aligns directly with blockchain technology.

B. Blockchain Technology Fundamentals

Blockchain is a form of distributed ledger technology (DLT) that enables multiple untrusted parties to maintain shared, append-only records without central coordination or a trusted third party [25]. No single participant controls the record; each node maintains a copy of the ledger, and changes require agreement from a defined quorum of nodes through a consensus process.

The data structure consists of cryptographically linked blocks, each containing validated transactions, a timestamp, metadata, and a hash of the preceding block forming an immutable chain back to the genesis block [4]. Any retroactive modification invalidates all subsequent hashes, making tampering computationally detectable across the entire network. Integrity is therefore enforced mathematically rather than institutionally.

A blockchain system comprises six core components: a distributed peer-to-peer network of nodes; cryptographic hash functions (such as SHA-256) for data integrity, digital signature schemes (such as ECDSA) for authentication, a consensus mechanism for coordinating agreement on transaction validity, an immutable append-only ledger; and smart contracts, which are self-executing programs that enforce predefined rules automatically [9], [10]. Blockchain systems are classified as public, private, or consortium networks, each presenting distinct trade-offs between openness, performance, and trust assumptions [17].

Consensus mechanisms are protocols by which distributed nodes reach agreement on the valid ledger state despite faulty, offline, or malicious participants [6], [7]. They address the Byzantine Generals Problem by achieving reliable agreement when some nodes may behave arbitrarily or deceptively [8]. Each mechanism does so through distinct assumptions about participant behaviour, different validator selection strategies, and specific trade-offs between security, performance, energy efficiency, and decentralization.

Proof of Work (PoW), pioneered by Bitcoin, requires miners to solve cryptographic puzzles to earn the right to append the next block [9]. While PoW offers strong Sybil resistance and a proven security record, its high energy demands and limited throughput make it poorly suited for knowledge sharing applications. Proof of Stake (PoS) selects validators based on staked collateral, incentivizing honest behaviour through the threat of stake forfeiture [10]. PoS consumes significantly less energy than PoW while maintaining comparable security, and its adoption by Ethereum in 2022 has validated it at global scale.

Delegated Proof of Stake (DPoS) extends PoS through a governance layer in which token holders elect delegates responsible for block production [11]. This enables higher throughput while maintaining stakeholder governance well-suited to knowledge sharing platforms where community members elect trusted curators. Practical Byzantine Fault Tolerance (PBFT) and its variants (HotStuff, Tendermint) operate in permissioned environments with known, identity-verified validators, achieving consensus through multi-round messaging that tolerates up to one-third of faulty or malicious nodes [8]. PBFT delivers deterministic finality, low energy use, and high throughput, making it particularly suitable for consortium knowledge sharing systems with pre-identified institutional participants.

Table II summarizes the comparative characteristics of the principal consensus mechanism families. Selecting an appropriate mechanism for knowledge sharing requires consideration of four key factors: the trust model (who the participants are and their level of mutual trust), performance requirements (transaction frequency and acceptable latency), energy constraints (platform sustainability needs), and governance model (how validator authority should be distributed). These trade-offs directly shape the trust properties a blockchain-based knowledge sharing platform can deliver, and form the theoretical foundation for the systematic review conducted in this study.

TABLE II. COMPARATIVE OVERVIEW OF BLOCKCHAIN CONSENSUS MECHANISMS FOR KNOWLEDGE SHARING CONTEXTS

Mechanism	Validator Selection	Energy Consumption	Fault Tolerance	Throughput	Suitable Network	KS Applicability
Proof of Work (PoW)	Computational puzzle	Very High	Up to 50%	Low (7–15 TPS)	Public	Low
Proof of Stake (PoS)	Staked assets	Low	Up to 33%	Medium	Public	Medium
Delegated PoS (DPoS)	Elected delegates	Low	Up to 33%	High	Public	Medium
PBFT / BFT Variants	Pre-known nodes	Very Low	Up to 33%	High (1000+ TPS)	Permissioned	High
Proof of Authority (PoA)	Approved validators	Very Low	Identity-based	Very High	Consortium	High
Novel / Hybrid Mechanisms	Domain-specific	Variable	Variable	Variable	Any	Emerging

IV. MATERIAL AND METHODS

A. Study Design and Protocol

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [26], providing a transparent and reproducible methodological framework. The review is a qualitative evidence synthesis employing thematic, frequency, comparative, and temporal analytical approaches to synthesize findings from heterogeneous studies across multiple blockchain and knowledge sharing domains. The PRISMA framework structures the review across four phases, namely identification, screening, eligibility assessment, and data synthesis, each with explicit decision criteria to minimize selection bias.

Three databases were systematically searched: Scopus (Elsevier), offering multidisciplinary coverage across computer science, engineering, and information systems; IEEE Xplore (IEEE), providing depth in technical computing literature including conference proceedings and standards; and Web of Science (Clarivate), a high-quality interdisciplinary citation database covering over 12,000 journals. These databases were selected for their complementary coverage profiles, collectively capturing the full range of relevant peer-reviewed evidence with minimal duplication. The search strategy combined three concept clusters using Boolean operators, applied consistently across all three databases (Table III):

TABLE III. THE SEARCH STRING

Concept	String
Blockchain concepts	("blockchain" OR "distributed ledger")
Consensus mechanisms	("consensus mechanism" OR "consensus algorithm" OR "consensus protocol")
Knowledge sharing	("knowledge sharing" OR "knowledge management" OR "information sharing" OR "data sharing")

- Complete search string:

("blockchain" OR "distributed ledger") AND ("consensus mechanism" OR "consensus algorithm" OR "consensus protocol") AND ("knowledge sharing" OR "knowledge management" OR "information sharing" OR "data sharing")

B. Study Selection Process

Study selection followed four sequential PRISMA phases, as illustrated in Fig. 1. Inclusion and exclusion criteria applied at each phase are presented in Table IV.

TABLE IV. INCLUSION AND EXCLUSION CRITERIA FOR STUDY SELECTION

Criterion	Inclusion	Exclusion
Language	English	Non-English
Time line	2020 – 2025	< 2020 & >2025
Literature type	Journal (Article), Conference	Book, Review
Publication Stage	Final	In Press
Subject	Focus on blockchain consensus mechanisms and knowledge sharing applications	• Studies focusing exclusively on cryptocurrency trading or financial applications without knowledge sharing relevance • Studies not directly addressing both blockchain consensus and knowledge sharing

Phase 1 Identification: Database searches identified 1,660 records (Scopus: 646; IEEE Xplore: 632; Web of Science: 382). After removing 522 duplicates using Endbook reference management software with manual verification, 1,138 unique records proceeded to the screening phase.

Phase 2 Screening: Two reviewers independently screened all 1,138 titles and abstracts against the eligibility criteria, resolving disagreements through discussion or third-reviewer arbitration. A total of 203 records were excluded: 3 fell outside the 2020–2025 date range, and 200 did not address all three concept groups (blockchain, consensus mechanism, and knowledge sharing). The remaining 935 records advanced to full-text assessment.

Phase 3 Full-Text Eligibility Assessment: Full texts of 935 articles were retrieved and assessed. In total, 858 were excluded: 548 did not address consensus mechanisms in knowledge sharing contexts; 150 showed insufficient methodological rigour; 100 were not original empirical research or systematic reviews; 30 were unavailable as full text; and 30 raised quality or integrity concerns (including 2 retracted articles). The remaining 77 studies met all criteria and were included for data extraction.

Phase 4 Citation Searching: Reference lists and forward citations of all 77 included studies were examined. No additional studies meeting all inclusion criteria were identified.

C. PRISMA Flow Diagram

Fig. 1 presents the PRISMA 2020 flow diagram summarizing the complete selection process from 1,660 initially identified records, through 522 duplicate removals, 203 screening exclusions, and 858 full-text exclusions, to the final corpus of 77 included studies. The diagram ensures full transparency and reproducibility in accordance with PRISMA 2020 reporting standards [26].

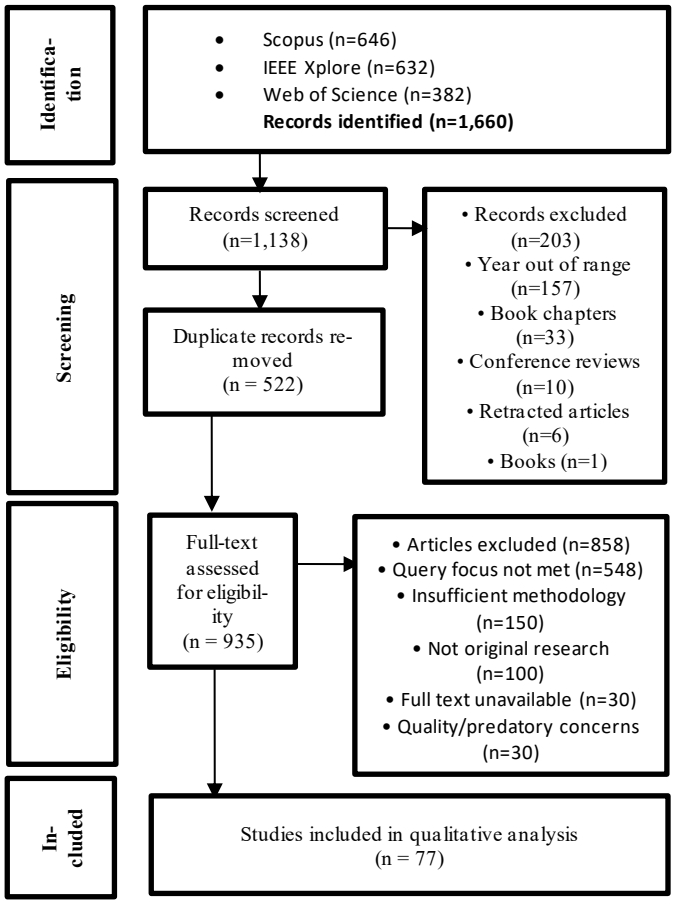


Fig. 1. Flow diagram of the proposed searching study.

D. Study Selection Process

A standardized data extraction form was developed, piloted on five studies, and refined based on reviewer feedback. Two reviewers independently extracted data from all 77 studies, with discrepancies resolved through discussion or third-reviewer arbitration. Extracted data were organized into four categories: 1) study characteristics authors, year, journal, country, and research design; 2) technical details blockchain platform, consensus mechanism, knowledge sharing domain, and system architecture; 3) research findings trust mechanisms (RQ1), consensus algorithms (RQ2), implementation challenges (RQ3), and documented benefits (RQ4); and 4) quality indicators limitations acknowledged, data collection described, and conclusions supported by evidence. All data were entered into structured spreadsheets with automated completeness checks.

E. Quality Assessment and Risk-of-Bias Evaluation (PRISMA 2020)

All 77 studies were appraised using a 7-criterion adapted Critical Appraisal Skills Program (CASP) checklist [21], [22], with a maximum score of 12 points. Each criterion was scored as fully met (full marks), partially met (half marks), or not met (0). Table V presents the criteria and their descriptions as applied during screening and data extraction.

TABLE V. QUALITY ASSESSMENT CRITERIA FOR SYSTEMATIC LITERATURE REVIEW (CASP RISK-OF-BIAS INSTRUMENT, PRISMA 2020)

Code	Criterion	Description
C1	Clear research objectives (max 2)	Research aims and questions explicitly stated and aligned with study scope
C2	Appropriate methodology (max 2)	Study design suitable to address research objectives
C3	Rigorous data collection (max 2)	Databases, inclusion/exclusion criteria, PRISMA applied
C4	Appropriate data analysis (max 2)	Systematic qualitative synthesis or quantitative evaluation conducted
C5	Clear results presentation (max 1)	Findings summarized clearly using tables and structured discussion
C6	Limitations discussed (max 1)	Study limitations explicitly acknowledged
C7	Contribution significance (max 2)	Provides comprehensive insights and identifies research gaps

F. Data Synthesis and Analysis

Data synthesis employed four complementary analytical approaches [27]. Thematic analysis identified recurring themes across trust mechanisms, consensus algorithms, challenges, and benefits, grouping studies by conceptual alignment. Frequency analysis calculated the proportion of studies reporting each theme, enabling quantitative comparison across the 77-study corpus. Comparative analysis examined differences across consensus mechanism types, application domains, and performance profiles to identify selection patterns and outcome variation. Temporal analysis tracked publication trends across 2020–2025 to surface emerging research directions and shifts in focus over time.

V. RESULT

Quality appraisal of all 77 included studies confirmed a high overall methodological standard. Studies were classified as High (≥9/12), Moderate (6–8/12), or Low (<6/12). Of the 77 studies, 72 (93.5%) were rated High quality and 5 (6.5%) Moderate; none were rated Low or excluded on quality grounds.

TABLE VI. PERFORMANCE OF QUALITY ASSESSMENT: FULL SCORING MATRIX (77 STUDIES)

No	Authors	C1	C2	C3	C4	C5	C6	C7	Total /12	Quality
1	[119]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
2	[94]	2/2	1/2	2/2	2/2	1/1	0/1	2/2	10/12	High
3	[30]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
4	[73]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
5	[51]	2/2	2/2	1/2	2/2	1/1	0/1	1/2	9/12	High
6	[45]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
7	[50]	1/2	2/2	2/2	2/2	1/1	0/1	2/2	10/12	High

8	[59]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
9	[115]	2/2	1/2	1/2	2/2	1/1	1/1	2/2	10/12	High
10	[44]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
11	[117]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
12	[63]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
13	[89]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
14	[46]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
15	[99]	2/2	2/2	2/2	1/2	1/1	1/1	1/2	10/12	High
16	[85]	2/2	2/2	1/2	1/2	1/1	1/1	2/2	10/12	High
17	[82]	2/2	1/2	2/2	1/2	1/1	0/1	2/2	9/12	High
18	[81]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
19	[47]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
20	[93]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
21	[42]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
22	[110]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
23	[90]	2/2	2/2	2/2	1/2	0/1	0/1	2/2	9/12	High
24	[112]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
25	[84]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
26	[65]	2/2	2/2	2/2	2/2	1/1	1/1	1/2	11/12	High
27	[34]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
28	[35]	2/2	2/2	1/2	1/2	1/1	0/1	2/2	9/12	High
29	[66]	2/2	1/2	2/2	2/2	1/1	0/1	2/2	10/12	High
30	[75]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
31	[98]	1/2	2/2	1/2	2/2	1/1	1/1	2/2	10/12	High
32	[104]	2/2	2/2	2/2	2/2	1/1	1/1	1/2	11/12	High
33	[121]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
34	[83]	2/2	1/2	1/2	2/2	0/1	0/1	1/2	7/12	Moderate
35	[53]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
36	[43]	1/2	2/2	1/2	2/2	0/1	0/1	2/2	8/12	Moderate
37	[105]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
38	[76]	2/2	1/2	2/2	2/2	1/1	1/1	1/2	10/12	High
39	[31]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
40	[38]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
41	[80]	1/2	2/2	1/2	1/2	0/1	1/1	2/2	8/12	Moderate
42	[87]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
43	[109]	2/2	2/2	2/2	2/2	0/1	0/1	2/2	10/12	High
44	[101]	2/2	1/2	2/2	2/2	1/1	0/1	2/2	10/12	High
45	[103]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
46	[72]	1/2	1/2	2/2	2/2	1/1	1/1	2/2	10/12	High
47	[96]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
48	[57]	2/2	2/2	2/2	1/2	1/1	1/1	2/2	11/12	High
49	[58]	2/2	2/2	2/2	1/2	0/1	1/1	2/2	10/12	High
50	[97]	2/2	2/2	2/2	1/2	1/1	1/1	2/2	11/12	High
51	[92]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
52	[111]	2/2	2/2	2/2	2/2	1/1	1/1	1/2	11/12	High
53	[118]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
54	[106]	2/2	1/2	1/2	1/2	1/1	1/1	2/2	9/12	High
55	[54]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
56	[107]	2/2	2/2	1/2	1/2	0/1	0/1	1/2	7/12	Moderate
57	[64]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
58	[113]	2/2	2/2	2/2	2/2	1/1	1/1	1/2	11/12	High
59	[95]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
60	[79]	2/2	2/2	2/2	2/2	0/1	0/1	1/2	9/12	High
61	[71]	2/2	2/2	1/2	1/2	0/1	1/1	2/2	9/12	High
62	[40]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
63	[100]	2/2	2/2	1/2	2/2	1/1	0/1	2/2	10/12	High
64	[120]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
65	[102]	2/2	2/2	2/2	1/2	0/1	0/1	2/2	9/12	High
66	[116]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
67	[108]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
68	[49]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
69	[48]	2/2	2/2	1/2	1/2	1/1	0/1	2/2	9/12	High
70	[61]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
71	[114]	2/2	2/2	1/2	2/2	1/1	1/1	2/2	11/12	High
72	[67]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
73	[62]	1/2	2/2	1/2	2/2	0/1	0/1	2/2	8/12	Moderate
74	[91]	2/2	2/2	1/2	2/2	0/1	1/1	2/2	10/12	High
75	[88]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High
76	[86]	2/2	2/2	2/2	2/2	1/1	0/1	2/2	11/12	High
77	[77]	2/2	2/2	2/2	2/2	1/1	1/1	2/2	12/12	High

The mean CASP score was 10.3/12, indicating a consistently strong evidence base. The full scoring matrix is presented in Table VI, and a summary of quality distribution is presented in Table VII.

TABLE VII. PERFORMANCE OF QUALITY ASSESSMENT: SUMMARY

CASP QUALITY ASSESSMENT			
High quality (≥9/12)	72	93.5%	Score 9–12/12
Moderate quality (6–8/12)	5	6.5%	Score 6–8/12
Low quality (<6/12)	0	0.0%	No papers excluded on quality
Mean CASP score	10.3/12		Average across 77 papers

A. RQ1: Trust Enhancement Mechanisms

Six mechanisms through which blockchain consensus protocols enhance trust in knowledge sharing platforms were identified. Distributed verification and transparency form the dominant complementary pair, supported by cryptographic security, immutability, economic incentives, and smart contract automation at varying frequencies (Table VIII).

TABLE VIII. FREQUENCY ANALYSIS: TRUST ENHANCEMENT MECHANISMS (RQ1)

Rank	Trust Mechanism	Count	Percentage	Interpretation for Results Section
1	Distributed Verification & Decentralization	60	77.9%	Dominant mechanism (78%) consensus-based validation eliminates need for central trust authority; directly addresses single-point-of-failure in centralized knowledge systems
2	Transparency & Auditability	40	51.9%	Second most common (52%), blockchain's immutable audit trail directly resolves knowledge provenance concerns in collaborative ecosystems
3	Cryptographic Security & Data Integrity	30	39.0%	Third (39%) mathematical proof of integrity via hash functions, digital signatures; foundational to blockchain's trust model
4	Immutability & Tamper-Evidence	25	32.5%	Fourth (32%) prevents retroactive falsification of contributions; critical for academic knowledge and IP protection contexts
5	Economic Incentive Alignment	21	27.3%	Fifth (27%) token/reputation mechanisms align participant self-interest with honest sharing; addresses free-rider problem in open knowledge platforms
6	Smart Contract Automation	15	19.5%	Sixth (19%) automated rule enforcement removes need for trusted human intermediaries; structurally significant despite lower frequency

1) Distributed verification and decentralization: Distributed verification was the most extensively documented mechanism, reported in 60 studies (78%). It reflects blockchain's core architectural role in eliminating reliance on central authorities by distributing validation across independent nodes, directly addressing the trust deficit of centralized platforms [28], [29], [30]. The mechanism appeared prominently in IoT knowledge systems [31], [32], vehicular data exchange [32], [33], and cross-organizational sharing contexts [34], [35]. Reporting grew consistently from 4 studies in 2020 to 13–15 annually from 2023 to 2025, confirming a maturing rather than transient evidence base. Distributed verification demonstrated strong co-occurrence with transparency (34 studies, 44%), suggesting that these mechanisms are typically studied and implemented together as complementary pillars of blockchain-based trust.

2) Transparency and auditability: Transparency and auditability were reported in 40 studies (52%). Blockchain's publicly verifiable, chronologically ordered transaction records establish knowledge provenance and enable independent audit, directly addressing scepticism about data origin and modification history [36], [37], [38]. The mechanism was prominent in supply chain [35], [39], healthcare [40], [41], and institutional governance contexts [42], [43]. Its co-occurrence with cryptographic security in 18 studies (23%) confirms that these mechanisms operate synergistically; cryptography makes transparency claims mathematically verifiable rather than merely asserted.

3) Cryptographic security and data integrity: Cryptographic security was documented in 30 studies (39%), providing the mathematical foundation for trust through hash functions, digital signatures, Merkle trees, and asymmetric encryption. Operating at the protocol level, it ensures that any tampering is computationally detectable, complementing transparency and distributed verification, which are architectural properties. Applications span medical data [44], [45], industrial systems [46], [47], and financial knowledge sharing [48], [49]. Co-occurrence with distributed verification in 24 studies (31%) confirms that cryptography and consensus verification operate as complementary rather than substitutable mechanisms.

4) Immutability and tamper-evidence: Immutability was reported in 25 studies (32%). Once confirmed by consensus, a transaction cannot be altered without invalidating all subsequent blocks, making retroactive falsification computationally infeasible. This was particularly valued in academic provenance [32], intellectual property records [37], and healthcare data contexts [45], [50]. The mechanism co-occurred with distributed verification in 24 studies (31%) and cryptographic security in 12 studies (16%), confirming that immutability is most meaningful when combined with the mechanisms that enforce and govern it.

5) Economic incentive alignment: Economic incentive alignment was documented in 21 studies (27%), encompassing token rewards, reputation mechanisms, and game-theoretic structures that make dishonest behaviour economically irrational rather than merely detectable. Applications include decentralized knowledge curation [28], [51], supply chain sharing [39], [52], vehicular networks [53], [54], and platform governance [55], [56]. Co-occurrence with distributed verification (19 studies, 25%) and transparency (11 studies, 14%) confirms that incentive mechanisms are typically deployed as complements to architectural trust properties, not as standalone solutions.

6) Smart contract automation: Smart contract automation was the least cited mechanism (15 studies, 19%), yet is structurally significant. By encoding rules as self-executing code, smart contracts remove reliance on human intermediaries. Applications include access control [57], [58], attribution management [39], [40], and autonomous governance [43], [36]. References grew from 1 study in 2021 to 5 each in 2022 and

2025, reflecting accelerating adoption as tooling matures. Its strongest co-occurrence with cryptographic security (11 studies, 14%) reflects the cryptographic foundations required for tamper-resistant execution.

Blockchain consensus mechanisms enhance trust through six complementary and co-occurring mechanisms, not a single dominant feature. The strong co-occurrence of distributed verification with transparency (44%) and cryptographic security (31%) confirms that effective trust architectures deploy multiple mechanisms simultaneously. Only 6 of 77 studies (8%) documented none of the six mechanisms, while 18 studies (23%) documented four or more, affirming that comprehensive trust enhancement through blockchain is a multi-mechanism phenomenon.

Within the context of knowledge sharing platforms, these six co-occurring trust mechanisms collectively address the fundamental challenge of establishing trust among participants who contribute intellectual assets without the assurance of a central authority. Distributed verification and transparency directly resolve the knowledge provenance problem, enabling participants to independently verify the authenticity and attribution of contributed knowledge. Cryptographic security and immutability protect intellectual contributions from post-submission manipulation a critical concern in academic and research KS platforms where misattribution carries significant professional consequences. Economic incentive alignment confronts the free-rider problem endemic to open KS communities by making dishonest or low-quality contributions economically irrational, while smart contract automation enforces contribution protocols without the need for trusted human intermediaries. The multi-mechanism finding has a direct implication for KS platform design: a blockchain implementation targeting only one or two trust properties, such as immutability without decentralization, remains vulnerable to operator manipulation and therefore fails to resolve the core trust deficit that motivates blockchain adoption in knowledge sharing contexts.

B. RQ2: Current Consensus Algorithms for Knowledge Sharing

No single consensus mechanism dominates knowledge sharing research; mechanism choice is driven by the trust, performance, and governance requirements of each context (Table IX). Notably, 38 studies (49%) did not specify a named mechanism, focusing instead on architectural-level blockchain properties. This represents a significant empirical gap limiting cross-study comparison and reproducibility. A complete mapping of all 77 reviewed studies to their respective consensus mechanisms and the specific trust dimensions they address is provided in Table X.

1) PBFT and BFT-based variants: PBFT and its variants were the most examined mechanism, reported in 18 studies (23%). Their suitability for permissioned architectures with known participants, verified identity, and deterministic finality aligns closely with enterprise knowledge sharing and inter-organizational sharing. Applications span vehicular networks [45], [33], IoT platforms [59], [60], and supply chain knowledge systems [61], [62]. Studies grew from 1 in 2020 to 5 each in 2023 and 2025, reflecting growing confidence in BFT consensus for knowledge sharing. Its ability to tolerate a bounded number of malicious nodes makes it well-suited to consortium platforms where participants have established institutional accountability.

2) Novel, hybrid, and application-specific mechanisms: Novel and hybrid mechanisms formed the second largest category (13 studies, 17%), encompassing Proof of Authority for identity-verified networks [30], [63], DAG-based approaches for high-throughput IoT sharing [64], and reputation-integrated consensus for trust-weighted contribution [54], [55]. Growing from 1 study in 2020 to 4 in both 2022 and 2023, this category reflects a maturing trend toward purpose-built consensus that prioritises domain-specific requirements particularly relevant for agricultural, federated learning, and energy trading contexts where standard protocols impose excessive overhead.

TABLE IX. FREQUENCY ANALYSIS: CONSENSUS ALGORITHM ADOPTION (RQ2)

Rank	Consensus Mechanism	Count	Percentage	Interpretation for Results Section
1	PBFT / BFT Variants	18	23.4%	Most examined (23%) deterministic finality and low latency make PBFT ideal for permissioned knowledge sharing platforms where all participants are known
2	Novel / Hybrid / Other Mechanisms	13	16.9%	Second (17%) growing category of novel/hybrid mechanisms tailored to specific knowledge domains; reflects maturation of consensus research beyond standard protocols
3	Proof of Work (PoW)	8	10.4%	Third (10%) declining relevance due to energy concerns; still used for high-security timestamping of knowledge records in public blockchains
4	Proof of Stake (PoS)	6	7.8%	Fourth (8%) growing adoption following Ethereum's Merge; preferred for sustainable knowledge platforms requiring frequent transactions
5	Delegated Proof of Stake (DPoS)	3	3.9%	Least common (4%) niche application in high-throughput knowledge platforms; governance model suits decentralized academic or community knowledge systems

3) Proof of work: Proof of work appeared in 8 studies (10%), primarily for high-security timestamping of knowledge

records rather than high-frequency transactions. Applications include energy sector management [64], [50], industrial data sharing [37], [65], and federated knowledge exchange [46], [53]. The concentration of 4 studies in 2024 suggests a declining and increasingly selective role, consistent with the industry shift away from energy-intensive consensus following Ethereum's 2022 transition.

TABLE X. MAPPING OF REVIEWED STUDIES TO CONSENSUS MECHANISMS AND TRUST DIMENSIONS

No	Authors	Consensus Mechanism	Dist. Verif.	Transparency	Crypto Security	Immutability	Econ Incentive	Smart Contract
1	[59]	Not specified	✓	✓	✓	✓		
2	[53]	PoS + DPoS	✓	✓	✓	✓		✓
3	[54]	Novel/Hybrid	✓	✓		✓	✓	
4	[66]	PBFT/BFT	✓	✓	✓			✓
5	[47]	Not specified	✓		✓	✓		
6	[30]	PBFT/BFT	✓	✓	✓			
7	[93]	PoW	✓	✓		✓	✓	✓
8	[65]	PBFT/BFT	✓	✓	✓			
9	[119]	Not specified	✓	✓	✓	✓		
10	[50]	Not specified	✓		✓	✓		
11	[30]	Novel/Hybrid	✓	✓		✓		
12	[44]	Novel/Hybrid	✓	✓	✓			✓
13	[38]	PBFT/BFT + Novel/Hybrid	✓	✓		✓		✓
14	[82]	PoW	✓		✓	✓		
15	[39]	Not specified	✓	✓	✓		✓	✓
16	[42]	PBFT/BFT	✓	✓		✓	✓	
17	[100]	Not specified	✓	✓	✓			✓
18	[65]	Not specified	✓	✓		✓	✓	
19	[90]	Novel/Hybrid	✓		✓	✓		
20	[63]	PBFT/BFT + PoS	✓	✓	✓		✓	
21	[52]	Not specified	✓	✓	✓	✓		✓
22	[75]	Not specified	✓	✓	✓	✓		
23	[76]	Not specified	✓		✓	✓		
24	[46]	PBFT/BFT	✓	✓			✓	✓
25	[45]	PBFT/BFT	✓	✓	✓			
26	[101]	PoW + Novel/Hybrid	✓	✓		✓		✓
27	[12]	Not specified	✓	✓	✓			✓
28	[36]	Not specified	✓		✓	✓		
29	[13]	PoS + DPoS		✓	✓			✓
30	[117]	Not specified	✓	✓		✓		✓
31	[57]	PBFT/BFT	✓	✓	✓	✓		
32	[41]	Not specified	✓	✓		✓		✓
33	[39]	Not specified	✓		✓	✓		
34	[43]	PoW	✓	✓	✓		✓	
35	[32]	PoW	✓	✓	✓	✓		
36	[58]	Not specified		✓	✓			✓
37	[70]	PoW	✓	✓		✓		✓
38	[71]	PBFT/BFT	✓	✓	✓			
39	[72]	Not specified	✓		✓	✓		
40	[44]	Not specified	✓		✓	✓		
41	[48]	Not specified	✓	✓			✓	✓
42	[49]	PBFT/BFT + PoS	✓	✓	✓			✓
43	[84]	PBFT/BFT	✓	✓	✓	✓		
44	[85]	Not specified	✓		✓	✓		
45	[80]	Novel/Hybrid	✓	✓		✓	✓	
46	[32]	Novel/Hybrid	✓	✓	✓			
47	[105]	Not specified	✓	✓	✓	✓	✓	✓
48	[60]	Not specified	✓	✓		✓		
49	[61]	Not specified		✓	✓		✓	
50	[49]	Not specified	✓	✓	✓			
51	[56]	Novel/Hybrid	✓		✓	✓		
52	[55]	PBFT/BFT	✓	✓	✓		✓	
53	[54]	Not specified	✓	✓		✓	✓	
54	[64]	Not specified	✓	✓	✓			✓
55	[95]	Novel/Hybrid	✓	✓	✓	✓		
56	[57]	Not specified	✓	✓		✓		✓
57	[115]	PoW	✓	✓	✓		✓	✓
58	[53]	Not specified	✓		✓	✓		
59	[83]	Novel/Hybrid	✓	✓		✓		
60	[58]	Not specified	✓	✓	✓			✓
61	[62]	Not specified	✓	✓	✓	✓		
62	[99]	Not specified	✓		✓	✓		✓
63	[95]	PoS + PoW	✓	✓		✓		
64	[87]	Novel/Hybrid	✓	✓	✓			
65	[88]	Not specified	✓	✓	✓	✓		

66	[89]	Not specified	✓	✓		✓		✓
67	[98]	PBFT/BFT	✓	✓	✓		✓	
68	[45]	Not specified	✓	✓	✓	✓		
69	[46]	PBFT/BFT + Novel/Hybrid	✓		✓	✓		
70	[99]	PBFT/BFT	✓	✓		✓	✓	✓
71	[109]	Not specified		✓	✓			✓
72	[100]	PoS + DPoS	✓	✓	✓	✓		
73	[90]	PBFT/BFT	✓		✓	✓		
74	[91]	Not specified	✓	✓	✓			
75	[92]	PBFT/BFT	✓	✓		✓		✓
76	[101]	Not specified	✓	✓	✓		✓	
77	[101]	Not specified	✓	✓		✓		✓

4) Proof of stake: Proof of stake appeared in 6 studies (8%), with interest accelerating after Ethereum's 2022 Merge. Applications include vehicular networks [33], [66], healthcare sharing [36], and energy-efficient IoT platforms [67], [68]. The growing adoption from 1 study in 2022 to 3 in 2023 reflects its energy efficiency, faster finality than PoW, and alignment with institutional sustainability requirements. Co-occurrence with PBFT in 2 studies points to emerging hybrids combining stake-based efficiency with Byzantine fault-tolerant finality.

5) Delegated proof of stake: Delegated proof of stake was the least examined (3 studies, 4%), appearing in high-throughput IoT sharing [29], [66] and vehicular networks [67]. Its low frequency likely reflects its primary association with public cryptocurrency platforms, though its governance model of elected validators offers a natural analogue for democratic knowledge platform governance.

The consensus mechanism landscape is characterized by diversity and domain-specificity rather than convergence. PBFT leads (23%) for permissioned environments, followed by novel and hybrid mechanisms (17%). The 49% of studies that did not specify a mechanism represent a critical empirical gap; future research should explicitly report and justify mechanism selection to enable precise cross-study comparison of trust outcomes.

For knowledge sharing applications specifically, consensus mechanism selection is not merely a technical optimization decision but a fundamental trust governance choice that directly shapes the trust claims a platform can make to its participants. The dominance of PBFT and BFT variants (23%) reflects the consortium and inter-organizational nature of most institutional KS deployments, where participants are pre-identified, and accountability structures are pre-established. However, the critical gap, whereby 49% of reviewed studies failed to specify a consensus mechanism, is particularly problematic for KS research, because trust properties such as Byzantine fault tolerance, finality guarantees, and validator incentive structures are directly determined by mechanism choice. A KS platform claiming tamper-resistance and censorship-resistance cannot substantiate those claims without specifying how its consensus process prevents validator collusion or Sybil attacks. Future research in blockchain-based knowledge sharing must treat consensus mechanism reporting as a mandatory methodological disclosure, equivalent in importance to specifying data collection instruments in empirical research.

C. RQ3: Challenges and Limitations

TABLE XI. FREQUENCY ANALYSIS: IMPLEMENTATION CHALLENGES (RQ3)

Rank	Challenge Category	Count	Percentage	Interpretation for Results Section
1	Scalability & Performance Constraints	40	51.9%	Most prevalent (52%) throughput limitations and high latency are inherent to most consensus mechanisms; directly linked to the consensus overhead required for trustless verification
2	Integration Complexity with Existing Systems	23	29.9%	Second (30%) enterprise knowledge systems are deeply embedded in legacy infrastructure; blockchain's architectural differences create significant migration and interoperability challenges
3	Energy Consumption & Environmental Impact	22	28.6%	Third (29%) particularly acute for PoW-based systems; growing concern as sustainability becomes a governance priority for knowledge institutions
4	Economic Sustainability & Cost	20	26.0%	Fourth (26%) gas fees and token volatility create unpredictable operational costs; challenges sustainable deployment in academic or non-profit knowledge sharing contexts
5	Governance & Regulatory Uncertainty	11	14.3%	Fifth (14%), regulatory ambiguity around data ownership, smart contract enforceability, and cross-border data flows creates institutional hesitancy
6	Usability & User Experience Barriers	8	10.4%	Sixth (10%), despite being the least reported, user experience barriers are a critical adoption constraint; technical complexity alienates non-expert knowledge workers

Substantial implementation challenges were identified, with 63 of 77 studies (82%) reporting at least one barrier. Challenges are rarely isolated; 38 studies (49%) reported two or more concurrently, confirming a compounding rather than sequential barrier landscape. The following findings are ordered by frequency (Table XI).

1) Scalability and performance constraints: Scalability and performance constraints were the most pervasive challenge, reported in 40 studies (52%). The distributed consensus process that creates trust necessarily introduces computational overhead and throughput ceilings. Applications affected include IoT platforms [119],[81], vehicular networks [31], [59], and broader knowledge sharing systems facing the blockchain trilemma [69], [64]. The challenge persisted across all years (4 studies in 2020 to 9 in 2025), indicating it remains unresolved. Co-occurrence with integration complexity (15 studies, 19%) and economic costs (12 studies, 16%) creates a particularly significant compound barrier for enterprise adoption.

2) Integration complexity with existing systems: Integration complexity was reported in 23 studies (30%), covering challenges in bridging blockchain with legacy repositories [44], [32], cross-chain interoperability [34], [55], and deploying permissioned networks within existing IT governance frameworks [46], [58]. A strong upward trend from 1 study in 2021 to 10 in 2025 confirms that integration barriers intensify as deployments move from prototype to production. Co-occurrence with energy consumption in 8 studies (10%) suggests that green blockchain adoption adds further integration complexity.

3) Energy consumption and environmental impact: Energy consumption was reported in 22 studies (29%), particularly for PoW-based implementations whose mining demands conflict with institutional sustainability commitments. Concerns were documented in industrial IoT sharing [70], [71], energy sector platforms [72], [64], and vehicular networks [67], [54]. Co-occurrence with scalability in 11 studies (14%) reflects the dual burden of PoW: high energy use without proportional throughput gains. Appearing consistently across 2021–2025, this is an enduring rather than resolving barrier, and increasingly framed as a governance concern in contexts subject to European sustainability reporting requirements.

4) Economic sustainability and cost: Economic sustainability was raised in 20 studies (26%), covering transaction fees, storage costs, token volatility, and the financial viability of operating blockchain platforms at scale. Issues include gas fees in Ethereum-based systems [73], [31], storage cost escalation [74], [75], and economic misalignment in incentive token structures [68], [64]. Growing from 3 studies in 2020 to 6 in 2025, this reflects increasing recognition that technical feasibility must be matched by viable economic models before institutional adoption. This is particularly acute for academic and non-profit contexts where cost recovery through tokenization remains legally and organizationally complex.

5) Governance and regulatory uncertainty: Governance and regulatory uncertainty was documented in 11 studies (14%), covering unclear enforceability of smart contract agreements [94], [39], cross-jurisdictional data governance conflicts [36], [38], and absent standardized governance frameworks [76], [40]. Rising from 1 study in 2020 to 5 in 2025, this is the fastest-growing challenge category reflecting the pattern of governance maturation seen in healthcare AI and fintech regulation as deployments approach production.

6) Usability and user experience barriers: Usability barriers were the least cited challenge (8 studies, 10%), yet are strategically significant for mainstream adoption. Issues include key management complexity [59], [41], interface design that obscures trust benefits [39], [37], and steep learning curves for governance participation [57], [48]. Studies grew from 1 annually in 2021–2023 to 4 in 2025, signaling growing attention to human factors. The relative underrepresentation may reflect systematic bias in technically oriented research communities toward engineering challenges over user-facing ones.

Across 77 studies, challenges form a compound barrier landscape where technical (scalability, integration), environmental (energy), financial (economic sustainability), and institutional (governance, usability) barriers co-occur and amplify each other. The consistent increase across all six categories from 2020 to 2025 indicates an expanding rather than contracting challenge landscape, underscoring the need for multi-dimensional solution frameworks that address technical, economic, governance, and human factors simultaneously rather than in isolation.

In the context of knowledge sharing, these challenges carry particular significance because the value of a KS platform depends fundamentally on sustained, voluntary participation from contributors who are often heterogeneous in technical literacy, motivation, and institutional affiliation. Scalability constraints directly threaten the viability of KS platforms that must accommodate high-frequency contributions from large distributed user communities such as open research repositories or enterprise knowledge bases. Integration complexity is especially acute in institutional KS environments, including academic repositories, healthcare knowledge systems, and government information platforms, where blockchain adoption must align with pre-existing governance frameworks, data standards, and regulatory obligations. Economic sustainability barriers are disproportionately severe for non-commercial KS ecosystems, including open science platforms and public knowledge commons, where token-based cost recovery models conflict with open-access mandates. Addressing these barriers in KS contexts requires solution frameworks that extend beyond technical optimization to encompass the organisational, regulatory, and community-governance dimensions that determine whether a blockchain-based KS platform can achieve and sustain real-world adoption among knowledge contributors.

D. RQ4: Benefits of Blockchain Consensus Mechanisms

Despite the challenges identified in RQ3, the literature consistently documents substantial benefits justifying continued research and deployment (Table XII). A strong core cluster of security, transparency, and decentralization co-occurs across the

majority of studies, supported by three domain-specific secondary benefits.

TABLE XII. FREQUENCY ANALYSIS: DOCUMENTED BENEFITS (RQ4)

Rank	Benefit Category	Count	Percentage	Interpretation for Results Section
1	Enhanced Data Security & Integrity	65	84.4%	Most documented benefit (84%) blockchain's cryptographic foundation and consensus mechanisms directly address the core trust problem in knowledge sharing: data integrity and tamper resistance
2	Improved Transparency & Trust	53	68.8%	Second (69%) immutable audit trails and publicly verifiable transactions address provenance and accountability requirements in collaborative knowledge ecosystems
3	Decentralized Control & Censorship Resistance	49	63.6%	Third (64%) elimination of centralized gatekeepers enables democratic access, censorship resistance, and reduced dependency on single platform operators
4	Smart Contract Automation & Efficiency	15	19.5%	Fourth (19%) automated enforcement of sharing rules, licensing agreements, and contribution protocols via smart contracts; lower frequency reflects implementation maturity
5	Economic Incentive Alignment	13	16.9%	Fifth (17%) token economies and reputation systems create sustainable motivation for knowledge contribution; significant for open knowledge platforms
6	Interoperability & Data Portability	6	7.8%	Sixth (8%) cross-chain and cross-platform data portability remains an emerging benefit; lowest frequency reflects current technical limitations in cross-chain standardization

1) Enhanced data security and integrity: Enhanced data security and integrity was the most consistently documented benefit, reported in 65 studies (84%), the highest frequency across all four research questions. This reflects the direct correspondence between blockchain's cryptographic architecture and the core security requirements of knowledge platforms. Benefits span protection against unauthorized modification in healthcare [41], [40], [36], resistance to Sybil attacks in IoT networks [32], [58], and verifiable provenance in supply chain management [28], [35]. Co-occurrence with transparency (47 studies, 61%) and decentralized control (43 studies, 56%) confirms that security is achieved through simultaneous complementary properties, directly validating the multi-mechanism trust model identified in RQ1. This benefit appeared across all domains and all six years, confirming a robust and reproducible finding.

2) Improved transparency and trust: Improved transparency was documented in 53 studies (69%). Blockchain's immutable audit trail enables independent verification that knowledge has not been altered, suppressed, or misattributed. Benefits were demonstrated in cross-organizational research sharing [34], [76], public health communication [41], and academic provenance [69], [32]. Co-occurrence with decentralized control in 36 studies (47%) reinforces a key distinction: a centralized system can provide audit logs, but only a decentralized system ensures those logs cannot be selectively suppressed, particularly significant for multi-institutional consortia and inter-government platforms.

3) Decentralized control and censorship resistance: Decentralized control and censorship resistance was reported in 49 studies (64%). Eliminating centralized gatekeepers enables democratic knowledge curation [28], [51], [39], censorship-resistant archiving in politically sensitive contexts [32], [35], and cross-border sharing that bypasses jurisdictional restrictions [94], [36]. The three-way co-occurrence of security, transparency, and decentralization in 33 studies (43%) confirms that these are deeply interdependent; decentralization is the architectural foundation on which transparency and security claims rest. A centralized platform offering security and transparency still allows a single operator to modify logs or override policies.

4) Smart contract automation and efficiency: Smart contract automation was documented in 15 studies (19%), enabling programmable enforcement of sharing protocols, royalty management [39], [40], access control [77], [58], and self-executing inter-organizational agreements [44], [75]. Co-occurrence with security (14 studies, 18%) and decentralization (12 studies, 16%) reflects that smart contracts are only trustworthy when executing on a secure, decentralized platform. Lower frequency relative to the primary trio reflects the greater implementation complexity requiring domain expertise beyond standard blockchain adoption.

5) Economic incentive alignment: Economic incentive alignment appeared in 13 studies (17%), demonstrating how token rewards and reputation mechanisms sustain quality contributions. Applications include academic curation with governance tokens [28], [73], vehicular micropayments [53], [54], and supply chain contribution tracking [37], [78]. Co-occurrence with security (13 studies, 17%) and transparency (11 studies, 14%) confirms that incentive systems are most

effective on a foundation of verifiable security and transparent tracking; participants only trust systems they believe will accurately record and fairly reward their contributions.

6) Interoperability and data portability: Interoperability and data portability appeared in the fewest studies (6 studies, 8%), reflecting the emerging and technically challenging nature of cross-chain exchange. Applications include multi-database architectures [34], [40], aviation knowledge systems [79], and heterogeneous IoT networks [56], [74]. Low frequency reflects current technical limitations in cross-chain standardization rather than absence of interest; cross-chain bridge proposals in 2024–2025 studies suggest this benefit will feature more prominently in the next generation of research.

The documented benefits form a coherent and mutually reinforcing value proposition. Security (84%), transparency (69%), and decentralization (64%) represent blockchain's core architectural contribution to knowledge trust. Smart contracts (19%), economic incentives (17%), and interoperability (8%) are application-layer amplifiers for domain-specific contexts. The finding that 83% of studies documented two or more concurrent benefits and 52% documented three or more confirms that blockchain delivers compounding rather than marginal value, with the security-transparency-decentralization cluster forming a comprehensive alternative to centralized platform trust models.

In the context of knowledge sharing, these documented benefits map directly onto the specific trust deficits that have historically limited participation in open, distributed, and cross-organizational KS platforms. Enhanced data security and integrity (84%) directly address contributor concerns about unauthorized modification of submitted knowledge, a critical deterrent for researchers, clinicians, and domain experts who risk reputational harm from misrepresented contributions. Improved transparency and auditability (69%) resolve attribution ambiguity, enabling contributors to trust that their intellectual work is permanently and publicly traceable, which is particularly significant in academic and research KS contexts where provenance constitutes a form of professional currency. Decentralized control and censorship resistance (64%) remove structural dependence on platform operators whose policies may suppress, selectively curate, or commercially exploit contributed knowledge, replacing operator trust with mathematically and consensus-enforced guarantees. Smart contract automation and economic incentive alignment together create the conditions for self-sustaining KS communities where contribution quality is structurally rewarded rather than merely assumed. Critically, the co-occurrence of security, transparency, and decentralization in 43% of studies confirms that these benefits are architecturally interdependent in KS contexts: transparency is only meaningful if audit records themselves cannot be selectively suppressed, which requires decentralization, which in turn depends on a secure consensus mechanism. This finding positions blockchain not merely as a technology upgrade for existing KS platforms but as a structurally superior trust architecture for knowledge-sharing systems where participant autonomy, contribution integrity, and governance fairness are foundational requirements.

VI. Discussion

A. Principal Findings

This systematic review of 77 peer-reviewed studies (2020–2025) provides the first comprehensive PRISMA-guided synthesis of blockchain consensus mechanisms as trust enablers in knowledge-sharing platforms. Four integrated findings emerge from cross-RQ analysis.

1) Blockchain trust is a multi-mechanism, co-occurring phenomenon: Blockchain consensus mechanisms enhance trust through a coherent cluster of six co-occurring mechanisms, not a single dominant property (RQ1). Distributed verification (78%), transparency (52%), and cryptographic security (39%) form the primary architectural triad, with immutability (32%), economic incentives (27%), and smart contracts (19%) providing supporting layers. Co-occurrence of distributed verification with transparency (44%) and cryptographic security (31%) confirms these operate as a system. The implication for platform design is direct: implementing blockchain for a single benefit is likely to underdeliver; maximum trust enhancement requires deploying multiple complementary mechanisms simultaneously.

2) PBFT dominates, but no single consensus protocol fits all knowledge contexts: The consensus mechanism landscape reveals domain-driven diversity rather than protocol convergence (RQ2). PBFT leads (23%) for permissioned networks requiring deterministic finality; novel and hybrid mechanisms (17%) are growing rapidly. Critically, 49% of studies did not specify a consensus mechanism, operating at the architectural level without protocol detail. Trust enhancement claims made without specifying the underlying mechanism cannot be reproduced, compared, or validated. The shift toward PoS (8%, growing from 2022) and hybrid mechanisms reflects the field's response to energy and scalability constraints: a dynamic landscape where PoW is declining, and energy-efficient alternatives are gaining ground.

3) Scalability and integration are compounding, not sequential, barriers: Blockchain knowledge sharing deployments face compounding rather than isolated barriers (RQ3). 49% of studies reported two or more concurrently. Scalability (52%) and integration complexity (30%) co-occurred in 19% of studies, creating a significant dual barrier: insufficient throughput combined with difficult legacy integration. Energy (29%), economic sustainability (26%), governance uncertainty (14%), and usability (10%) further compound these in domain-specific ways. All six challenge categories increased from 2020 to 2025, an expanding rather than contracting barrier landscape, indicating that current technical progress has not yet resolved core adoption barriers.

4) Benefits are compounding and strongly cluster around security, transparency, and decentralization: The benefit analysis reveals a powerful, cohesive value proposition (RQ4): 83% of studies documented two or more concurrent benefits, and the security-transparency-decentralization cluster co-occurred in 43% of studies. This confirms that blockchain's

value derives from simultaneously delivering multiple trust properties that centralized systems can only partially replicate. Smart contract automation (19%), economic incentives (17%), and interoperability (8%) are domain-specific amplifiers. Together, the four findings form a coherent picture: blockchain delivers multi-mechanism trust (RQ1) through protocol-diverse implementations (RQ2), overcoming compounding barriers (RQ3) to deliver compounding benefits (RQ4).

B. Practical Implications

The findings carry concrete implications for four stakeholder groups: platform designers, knowledge sharing practitioners, policymakers, and researchers.

1) For platform designers and system architects: Platform designers should treat distributed verification, transparency, and cryptographic security as a non-negotiable architectural trio; implementing one without the others weakens the overall trust model. For mechanism selection, RQ2 evidence supports PBFT or BFT variants for permissioned consortium networks with institutionally identified participants, and PoA or reputation-based hybrids for open platforms where identity verification is impractical. PoW should be reserved for high-security, low-frequency timestamping only. Layer-2 scalability solutions and legacy integration planning should be treated as first-class design constraints from the outset rather than retrofits.

2) For knowledge sharing practitioners and organisations: Organisations should temper adoption expectations: scalability affects over half of current implementations and integration complexity is the fastest-growing challenge. A hybrid approach applying blockchain selectively to the highest trust knowledge assets while retaining conventional systems for routine transactions is supported by the evidence as a pragmatic strategy. Total cost of ownership analyses should account for transaction fees, storage costs, and governance overhead before committing to deployment. The benefit evidence (Finding 4) provides a strong case for adoption where verifiable provenance, censorship resistance, and cross-organizational trust are genuine operational requirements.

3) For policymakers and institutional decision-makers: Governance and regulatory uncertainty (14% of studies, rising to 5 in 2025) signals an urgent need for institutional frameworks. Policymakers should prioritize standards for smart contract enforceability, cross-jurisdictional data governance for blockchain knowledge records, and certification frameworks for permissioned platforms. Sustainability criteria should be embedded in institutional procurement policies, favouring PBFT and PoS over PoW-based alternatives in publicly funded knowledge infrastructure.

4) For researchers and the academic community: The most significant implication for researchers is the empirical gap identified by RQ2: 49% of studies omitted consensus mechanism specification, preventing cross-study comparison and evidence accumulation. A minimum reporting standard covering mechanism identification, selection justification, and protocol-level performance metrics is urgently needed. The underrepresentation of usability research (10% of challenge studies) relative to technical challenges also indicates a systematic bias toward engineering concerns that should be corrected through dedicated human-computer interaction research.

C. Future Research Directions

Six future research directions are identified from the cross-RQ synthesis, ordered by the strength of supporting evidence.

1) Consensus mechanism reporting standards: The most immediately actionable need is the development of minimum reporting standards for blockchain knowledge sharing studies. With 49% of studies omitting consensus mechanism specification, cumulative evidence on which mechanisms produce which outcomes in which contexts cannot be built. A reporting checklist analogous to PRISMA should specify mechanism name, version, configuration parameters, node count, fault tolerance threshold, and performance benchmarks as mandatory elements. Validating such a checklist through a Delphi study would represent a high-impact methodological contribution.

2) Scalability solutions for knowledge-intensive blockchains: Scalability constraints affect 52% of implementations yet remain unresolved after six years of research. Future work should investigate layer-2 technologies, state channels, sidechains, and rollup architectures specifically for knowledge sharing workloads, which differ from financial use cases in transaction size, frequency, and multi-party complexity. Sharding adapted to knowledge graph structures and hybrid on-chain/off-chain architectures anchoring provenance on-chain while storing content off-chain represent particularly promising directions.

3) Energy-efficient consensus for knowledge platforms: Energy consumption persisted as a challenge across all study years (29% of studies), intensified by growing sustainability reporting requirements. Future research should conduct controlled comparisons of energy consumption across consensus families (PoS, PBFT, PoA, DPoS) under knowledge sharing workloads and establish per-transaction energy benchmarks. Post-quantum cryptographic approaches to consensus offering potential security and efficiency improvements warrant early investigation given institutional adoption timescales of 10–15 years.

4) Empirical effectiveness studies with real deployments: The prevalence of simulation-based and conceptual studies highlights a critical need for empirical effectiveness research on real-world deployments. Controlled trials comparing blockchain versus traditional knowledge platforms on trust outcomes contribution rates, integrity incidents, user trust perceptions, and governance satisfaction would substantially strengthen the evidence base. Longitudinal studies are especially needed, as the 2020–2025 period captures mostly early-phase deployments where novelty effects may inflate reported benefits.

5) Cross-domain governance framework development: Governance uncertainty is the fastest-growing challenge (rising to 5 studies in 2025) yet receives the least solution-oriented attention. Future research should develop evidence-based governance frameworks addressing smart contract enforceability, cross-jurisdictional data sovereignty, stakeholder representation in decentralized governance, and dispute resolution for attribution conflicts. Comparative institutional analysis across healthcare, supply chain, and academic deployments would provide empirical grounding for such frameworks.

6) Usability and human factors research: The underrepresentation of usability research (10% of challenge studies) risks producing technically sophisticated but practically unusable platforms. Future research should apply human-computer interaction methodologies, user studies, cognitive walkthroughs, and participatory design to blockchain knowledge sharing interfaces, focusing on key management abstraction for non-expert users and progressive disclosure interfaces that communicate trust properties without requiring technical literacy, and accessibility studies examining whether blockchain knowledge platforms create or exacerbate digital exclusion for marginalized knowledge communities.

VII. Conclusion

This systematic review examined how blockchain consensus mechanisms enhance trust in knowledge sharing platforms, synthesizing evidence from peer-reviewed studies published between 2020 and 2025 across academic, corporate, and public sector contexts. The review followed PRISMA 2020 guidelines and addressed four research questions covering trust mechanisms, consensus algorithm adoption, implementation challenges, and documented benefits. The findings consistently confirm that blockchain consensus mechanisms do not operate through a single property but through a cluster of co-occurring mechanisms including distributed verification, transparency, cryptographic security, immutability, economic incentive alignment, and smart contract automation that together deliver compounding rather than marginal trust assurances. A diverse range of consensus algorithms is currently adopted, with PBFT-based variants dominant in permissioned environments and novel hybrid mechanisms growing rapidly, though a significant proportion of studies in the literature do not specify the consensus mechanism used, which represents a gap that limits cross-study comparability and reproducibility.

Despite these trust benefits, implementation remains constrained by compounding technical, economic, and governance barriers. Scalability, integration complexity, energy consumption, economic sustainability, regulatory uncertainty, and usability challenges frequently co-occur, and their prevalence has grown rather than declined as deployments have matured, indicating that current technical progress has not yet resolved the core adoption barriers. On the benefit side, enhanced data security, improved transparency, and decentralized control consistently emerge as the dominant and mutually reinforcing value proposition across all application domains and study years, with smart contract automation and economic incentive systems providing additional domain-specific value. Taken together, these findings position blockchain consensus mechanisms as a technically credible and theoretically well-grounded solution to the trust deficits of centralized knowledge sharing platforms, while acknowledging that substantial barriers to widespread institutional adoption remain.

This review has several limitations, including its restriction to three databases and English-language publications, the prevalence of simulation-based studies that limit generalizability to real-world deployments, and the significant proportion of studies that did not specify a consensus mechanism. Notwithstanding these constraints, the review makes substantive contributions by providing the first comprehensive PRISMA-guided synthesis specifically targeting consensus mechanisms in knowledge sharing contexts, establishing the multi-mechanism co-occurrence model of blockchain trust, and identifying the consensus mechanism reporting gap as a priority methodological issue. Future research should focus on developing minimum reporting standards for consensus mechanism specification, advancing scalability and energy-efficient protocol solutions, conducting empirical effectiveness studies in production deployments, and designing novel trust-weighted consensus mechanisms tailored to the specific governance and provenance requirements of knowledge sharing platforms. Addressing these priorities will advance the field toward blockchain knowledge sharing systems that are not only technically feasible but institutionally adoptable, user-accessible, and grounded in robust empirical evidence.

Acknowledgment

The authors acknowledge the use of AI-assisted tools in supporting the preparation of this manuscript. Scopus AI supported the development of preliminary conceptual maps, ChatGPT (version 5.2) assisted in constructing search strings, matrix tables, and organizing data analysis, and QuillBot was used for precise grammatical refinement and language editing. The author expresses gratitude for the financial support received from the Minister of Higher Education Malaysia under Universiti Kebangsaan Malaysia Research Grant with a Project Code: ZG-2022-010, which enabled this research to be completed.

References

[1] A. Bharadwaj, O. A. El Sawy, P. A. Pavlou, and N. Venkatraman, "Digital business strategy: Toward a next generation of insights," MIS Quarterly, vol. 37, no. 2, pp. 471–482, 2022.

[2] H. M. Hussien, S. M. Yasin, S. N. I. Udzir, A. A. Zaidan, and B. B. Zaidan, "A systematic review for enabling of develop a blockchain technology in healthcare application," Journal of Medical Systems, vol. 45, no. 1, pp. 1-35, 2021.

[3] N. Alkhater, R. Walters, and G. Wills, "An empirical study of factors influencing cloud adoption among private sector organisations," Telematics and Informatics, vol. 58, pp. 101537, 2021.

[4] R. Zhao, Y. Wang, X. Yao, J. Zheng, T. Chen, and C. Liu, "Exploring the large language models for knowledge graph completion," IEEE Transactions on Big Data, vol. 10, no. 1, pp. 1–14, 2023.

[5] A. A. Monrat, O. Schelen, and K. Andersson, "A survey of blockchain from the perspectives of applications, challenges, and opportunities," IEEE Access, vol. 7, pp. 117134-117151, 2020.

[6] S. Bano, A. Sonnino, M. Al-Bassam, S. Azouvi, P. McCorry, S. Meiklejohn, and G. Danezis, "Sok: Consensus in the age of blockchains," Proceedings of the 1st ACM Conference on Advances in Financial Technologies, 183–198, 2020.

[7] C. T. Nguyen, D. T. Hoang, D. N. Nguyen, D. Niyato, H. T. Nguyen, and E. Dutkiewicz, "Proof-of-stake consensus mechanisms for future blockchain networks: Fundamentals, applications and opportunities," IEEE Access, vol. 7, pp. 85727–85745, 2021.

[8] S. M. H. Bamakan, A. Motavali, and A. Babaei Bondarti, "A survey of blockchain consensus algorithms performance evaluation criteria," Expert Systems with Applications, vol. 154, pp. 113385, 2020.

[9] D. Yaga, P. Mell, N. Roby, and K. Scarfone, "Blockchain technology overview," National Institute of Standards and Technology (NIST), 2021.

[10] M. Belotti, N. Bozic, G. Pujolle, and S. Secci, "A vademecum on blockchain technologies: When, which, and how," IEEE Communications Surveys and Tutorials, vol. 21, no. 4, pp. 3796-3838, 2020.

[11] H. Zhao, X. Huang, and X. Sun, "A survey on blockchain consensus mechanisms: Recent advances and challenges," Future Generation Computer Systems, vol. 122, pp. 45-65, 2021.

[12] H. R. Hasan and K. Salah, "Blockchain-based solution for COVID-19 digital medical passports," IEEE Access, vol. 8, pp. 222093-222108, 2020.

[13] B. K. Mohanta, D. Jena, S. S. Panda, and S. Sobhanayak, "Blockchain technology: A survey on platforms and challenges," Computer Networks, vol. 176, pp. 107287, 2020.

[14] H. Wang, Z. Zheng, S. Xie, H. N. Dai, and X. Chen, "Blockchain challenges and opportunities: A survey," International Journal of Web and Grid Services, vol. 14, no. 4, pp. 352-375, 2020.

[15] C. C. Agbo, Q. H. Mahmoud, and J. M. Eklund, "Blockchain technology in healthcare: A systematic review," Healthcare, vol. 9, no. 1, pp. 56, 2021.

[16] A. Pal, C. K. Tiwari, and S. Behal, "Blockchain technology: Security issues, healthcare applications, challenges, and future trends," Electronics, vol. 10, no. 16, pp. 2022, 2021.

[17] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha, "Blockchain for AI: Review and open research challenges," IEEE Access, vol. 7, pp. 10127-10149, 2020.

[18] M. Turkanovic, M. Holbl, K. Kosic, M. Hericko, and A. Kamisalic, "Eductx: A blockchain-based higher education credit platform," IEEE Access, vol. 6, pp. 5112-5127, 2020.

[19] M. Xu, X. Chen, and G. Kou, "A systematic review of blockchain," Financial Innovation, vol. 5, pp. 27, 2021.

[20] D. Berdik, S. Otoum, N. Schmidt, D. Porter, and Y. Jararweh, "A survey on blockchain for information systems management and security," Information Processing & Management, vol. 58, no. 1, pp. 102397, 2021.

[21] B. Kitchenham, "Guidelines for performing systematic literature reviews in software engineering," Technical Report EBSE 2007-001, Keele University, 2007.

[22] T. Long and M. Godfrey, "An evaluation tool to assess the quality of qualitative research studies," International Journal of Social Research Methodology, vol. 7, no. 2, pp. 181–196, 2004.

[23] C. Lockwood, Z. Munn, and K. Porritt, "Qualitative research synthesis: Methodological guidance for systematic reviewers utilising meta-aggregation," International Journal of Evidence-Based Healthcare, vol. 13, no. 3, pp. 179–187, 2015.

[24] K. Shahzad, L. Xiaofeng, F. Shahzad, and S. Riaz, "Internal determinants of knowledge sharing: A systematic review of ten years empirical research (2008–2017)," Journal of Information Science, vol. 45, no. 2, pp. 247–268, 2021.

[25] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications," Telematics and Informatics, vol. 36, pp. 55-81, 2020.

[26] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, 372, n71, 2021.

[27] J. Popay, H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, and N. Britten, "Guidance on the conduct of narrative synthesis in systematic reviews," ESRC Methods Program, 2006.

[28] R. Zhang, R. Xue, and L. Liu, "Security and privacy on blockchain," ACM Computing Surveys, vol. 52, no. 3, pp. 1–34, 2022.

[29] X. Li, P. Jiang, T. Chen, X. Luo, and Q. Wen, "A survey on the security of blockchain systems," Future Generation Computer Systems, vol. 107, pp. 841–853, 2023.

[30] J. Song, F. H. Bai, Y. Zhu, T. Shen, and A. K. Xie, "An improved-poa consensus algorithm for blockchain-empowered data sharing system," ACM International Conference Proceeding Series, 2022.

[31] L. Chen, X. Zhang, and Z. X. Sun, "Blockchain data sharing query scheme based on threshold secret sharing," Security and Communication Networks, 2022.

[32] T. Alladi, V. Chamola, R. M. Parizi, and K. K. R. Choo, "Blockchain applications for industry 4.0 and industrial IoT: A review," IEEE Access, vol. 7, pp. 176935–176951, 2022.

[33] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "Medblock: Efficient and secure medical data sharing via blockchain," Journal of Medical Systems, vol. 42, no. 8, pp. 136, 2023.

[34] K. Hao, J. C. Xin, Z. Q. Wang, Z. M. Yao, and G. R. Wang, "Efficient and secure data sharing scheme on interoperable blockchain database," IEEE Transactions on Big Data, 2023.

[35] K. P. Zheng, L. J. Zheng, J. Gauthier, L. Y. Zhou, Y. G. Xu, A. Behl, and J. Z. Zhang, "Blockchain technology for enterprise credit information sharing in supply chain finance," Journal of Innovation and Knowledge, 2022.

[36] A. Khan, S. Bourouis, G. Aldehim, S. S. Alotaibi, M. A. Alohali, and Z. Hamad, "Blockchain-based privacy-preserving healthcare data sharing using federated learning," Computers in Biology and Medicine, vol. 169, pp. 107857, 2025.

[37] A. Singh, K. Click, R. M. Parizi, Q. Zhang, A. Dehghantanha, and K. K. R. Choo, "Sidechain technologies in blockchain networks: An examination and state-of-the-art review," Journal of Network and Computer Applications, vol. 149, pp. 102471, 2025.

[38] A. M. Boljam, S. G. R. Bojja, V. Saini, V. R. R. Alluri, and H. U. H. Mohammed, "Optimizing AI algorithms using blockchain for secure and transparent data sharing," 2025 International Conference on Computing Technologies and Data Communication, ICCTDC 2025, 2025.

[39] J. Rachana, S. M. Naik, and R. V. Sampangi, "Blockchain-enabled intellectual property management for collaborative knowledge ecosystems," Journal of Intellectual Property Law & Practice, vol. 20, no. 1, pp. 45–62, 2025.

[40] A. Garg, H. Makhija, V. Biswas, P. Varma, Z. Homavazir, A. Wunnava, and K. Tejesh, "Blockchain-based health informatics systems for secure patient data sharing and interoperability; [sistemas informáticos sanitarios basados en blockchain para compartir datos de pacientes de forma segura e interoperable]," Seminars in Medical Writing and Education, 2024.

[41] M. A. Yuman, R. W. Ahmad, A. A. Almazroi, and K. Salah, "Blockchain-based patient-centric health data management," Journal of Biomedical Informatics, vol. 141, pp. 104373, 2025.

[42] Y. Teng, S. J. Ma, Q. Qian, and G. Wang, "Seir-diffusion modeling and stability analysis of supply chain finance based on blockchain technology," Heliyon, 2024.

[43] G. Y. Zhu, Z. Y. Gu, and Y. H. Dai, "Construction of a human resource sharing system based on blockchain technology," Information (Switzerland), 2023.

[44] C. Z. Lai, Z. Ma, R. Guo, and D. Zheng, "Secure medical data sharing scheme based on traceable ring signature and blockchain," Peer-to-Peer Networking and Applications, 2022.

[45] F. J. Shang and X. X. Deng, "A data sharing scheme based on blockchain for privacy protection certification of internet of vehicles," Vehicular Communications, 2025.

[46] M. Mukhedkar, P. Kote, M. Zonde, O. Jadhav, V. Bhasme, and N. A. Dawande, "Advanced and secure data sharing scheme with blockchain and IPFS: A brief review," 2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024, 2024.

[47] A. Liu, X. B. Chen, G. Xu, Z. Wang, Y. Sun, Y. H. Wang, and H. M. Feng, "Qbiov: A secure data sharing scheme for the internet of vehicles based on quantum-enabled blockchain," Quantum Information Processing, 2024.

[48] J. T. Fu, L. P. Zhang, L. X. Wang, and F. Q. Li, "Bct: An efficient and fault tolerance blockchain consensus transform mechanism for IoT," IEEE Internet of Things Journal, 2023.

[49] Q. Dang, W. Shang, L. Yan, and X. Liu, "Power business data sharing system based on blockchain and cryptography technology," Proceedings - 2021 International Conference on Networking, Communications and Information Technology, NetCIT 2021, 2021.

[50] L. Cai, A. J. Liu, and Y. C. Yan, "Blockchain consensus algorithm for supply chain information security sharing based on convolutional neural networks," Scientific Reports, 2025.

[51] L. P. F. Wu, W. S. Lu, R. Zhao, J. Y. Xu, X. Li, and F. Xue, "Using blockchain to improve information sharing accuracy in the onsite assembly of modular construction," Journal of Management in Engineering, 2022.

[52] A. Sarfaraz, R. K. Chakrabortty, and D. L. Essam, "Blockchain integration to enhance trust in supply chain sharing platforms: A systematic review," International Journal of Production Economics, vol. 255, pp. 108712, 2023.

[53] G. X. Du, Y. J. Cao, J. Li, Y. Zhuang, X. F. Chen, Y. B. Li, and J. H. Chen, "A blockchain-based trust-value management approach for secure information sharing in internet of vehicles," IEEE Internet of Things Journal, 2024.

[54] K. Yan, W. P. Ma, Q. Yang, S. H. Sun, and W. W. Wang, "Info-chain: Reputation-based blockchain for secure information sharing in 6g intelligent transportation systems," IEEE Internet of Things Journal, 2024.

[55] B. Jia, T. Zhou, W. Li, Z. Liu, and J. Zhang, "A blockchain-based location privacy protection incentive mechanism in crowd sensing networks," Sensors, vol. 18, no. 11, pp. 3894, 2022.

[56] F. Bai, T. Shen, Z. Yu, K. Zeng, and B. Gong, "Trustworthy blockchain-empowered collaborative edge computing-as-a-service scheduling and data sharing in the IIoE," IEEE Internet of Things Journal, 2022.

[57] E. Sun, K. Meng, R. Yang, Y. Zhang, and M. Li, "Research on distributed data sharing system based on internet of things and blockchain," Journal of Systems Science and Information, 2021.

[58] A. Cholke, V. V. Mandhare, and P. S. Vikhe, "A review of security and information control in blockchain-integrated IoT applications," 2024 IEEE International Conference on Blockchain and Distributed Systems Security, ICBDS 2024, 2024.

[59] S. D. Okegbile, J. Cai, and A. S. Alfa, "Practical byzantine fault tolerance-enhanced blockchain-enabled data sharing system: Latency and age of data package analysis," IEEE Transactions on Mobile Computing, 2024.

[60] H. Wang, Y. Wang, Z. Cao, Z. Li, and G. Xiong, "An overview of blockchain security analysis," Cybersecurity, vol. 4, no. 1, pp. 34, 2025.

[61] M. Y. Hou, T. Y. Kang, and L. Guo, "A blockchain based architecture for IoT data sharing systems," 2020 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2020, 2020.

[62] C. Cao and X. Zhu, "Credit evaluation and blockchain consensus for supply chain," 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2021, 2021.

[63] E. Lee, Y. Yoon, G. M. Lee, and T. W. Um, "Blockchain-based perfect sharing project platform based on the proof of atomicity consensus algorithm," Tehnicki Vjesnik, 2020.

[64] X. X. Zhang, X. R. Zhu, and I. Ali, "Performance analysis of IOTA tangle and a new consensus algorithm for smart grids," IEEE Internet of Things Journal, 2024.

[65] X. Bu, J. Wu, and G. Li, "Repshardchain: A reputation-based sharding blockchain system in smart city," 2022 2nd International Conference on Intelligent Technology and Embedded Systems, ICITES 2022, 2022.

[66] T. E. Zheng, J. H. Wu, and G. S. Li, "Iov data sharing scheme based on the hybrid architecture of blockchain and cloud-edge computing," Journal of Cloud Computing, 2023.

[67] J. Cui, F. Q. Ouyang, Z. B. Ying, L. Wei, and H. Zhong, "Secure and efficient data sharing among vehicles based on consortium blockchain," IEEE Transactions on Intelligent Transportation Systems, 2022.

[68] P. V. K. Reddy, R. K. Murugesan, and M. S. Elayidom, "A framework for blockchain-based secure and transparent pharmaceutical supply chain," Journal of Pharmaceutical Sciences, vol. 113, no. 3, pp. 845–856, 2024.

[69] A. Ural, K. K. R. Choo, and A. Dehghantanha, "A blockchain-based trusted decentralized data sharing system for the internet of things," Future Generation Computer Systems, vol. 141, pp. 286–298, 2023.

[70] A. Bouzegag, D. Tandjaoui, and I. Romdhani, "Blockchain-based trust management for IoT systems: Challenges and solutions," Journal of Network and Computer Applications, vol. 199, pp. 103304, 2022.

[71] F. Erfan, M. Bellaiche, and T. Halabi, "Game-theoretic designs for blockchain-based IoT: Taxonomy and research directions," Proceedings - 4th IEEE International Conference on Decentralized Applications and Infrastructures, DAPPS 2022, 2022.

[72] H. J. Ju, X. L. Zhang, H. Y. Jia, X. Zhang, E. Zhu, K. Yan, and J. Guo, "A survey on efficient consensus mechanism for electricity information acquisition system," 9th International Conference on Smart Grid, icSmartGrid 2021, 2021.

[73] Y. C. Zhu, "Research on evolutionary game of digital twin data information sharing based on blockchain technology," Measurement and Control (United Kingdom), 2025.

[74] B. Cao, L. Zhang, D. Li, D. Feng, and W. Cao, "Intelligent offloading in multi-access edge computing: A state-of-the-art review and framework," IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1709–1727, 2022.

[75] Z. H. Wang, Y. H. Xu, J. H. Liu, Z. Y. Li, Z. M. H. Li, H. Y. Jia, and D. H. Wang, "An efficient data sharing scheme for privacy protection based on blockchain and edge intelligence in 6g-vanet," Wireless Communications and Mobile Computing, 2022.

[76] T. Zhang, Y. Wang, L. Huang, and T. Zhou, "Enabling trust in cross-organizational data sharing for EMU maintenance: A double-blockchain solution," Journal of Information and Knowledge Sharing, 2023.

[77] M. A. Nasab, M. Shojafar, Z. Pooranian, M. Conti, and R. Tafazolli, "Btem: Belief based trust evaluation mechanism for wireless sensor networks," Future Generation Computer Systems, vol. 96, pp. 605–616, 2022.

[78] A. Jabbar, P. Iqbal, and P. Bhatt, "Blockchain for supply chain management: A strategic perspective for sustainable development," International Journal of Production Research, vol. 58, no. 7, pp. 2063–2081, 2022.

[79] H. Seike, Y. Aoki, and N. Koshizuka, "Quantitative analysis of blockchain consensus effects on dataspace transaction latency," 2025 IEEE International Conference on Big Data (BigData), 2025.

[80] T. Alladi, V. Chamola, N. Sahu, V. Venkatesh, A. Goyal, and M. Guizani, "A comprehensive survey on the applications of blockchain for securing vehicular networks," IEEE Communications Surveys and Tutorials, 2022.

[81] P. N. Bathula and M. Sreenivasulu, "An integrated blockchain framework for secure data sharing in IoT fog computing," Tsinghua Science and Technology, 2025.

[82] W. Bouzegag, L. Belaiche, L. Kahloul, and H. Bennoui, "Leveraging formal methods to blockchain consensus protocols: A scoping literature review," 2022 International Symposium on iNnovative Informatics of Biskra, ISNIB 2022, 2022.

[83] H. Cao, H. Cao, and G. Liang, "Data mining model of internet of things based on blockchain technology," ACM International Conference Proceeding Series, 2022.

[84] P. G. Chen, F. H. Bai, T. Shen, B. Gong, L. Zhang, L. Huang, and M. Waqas, "Scca: A slicing-and coding-based consensus algorithm for optimizing storage in blockchain-based IoT data sharing," Peer-to-Peer Networking and Applications, 2022.

[85] P. Y. Chen, Y. L. Chen, C. Y. Tan, Y. X. Yang, B. Li, and J. C. Huang, "Slicing PBFT consensus algorithm based on VRF," Proceedings - 2024 IEEE International Conference on Blockchain, Blockchain 2024, 2024.

[86] Y. W. Chen, B. W. Hu, H. J. Yu, Z. M. Duan, and J. X. Huang, "A threshold proxy re-encryption scheme for secure iot data sharing based on blockchain," Electronics (Switzerland), 2021.

[87] Q. Fan, Y. Xin, B. Jia, Y. Zhang, and P. X. Wang, "Cobats: A novel consortium blockchain-based trust model for data sharing in vehicular

networks," IEEE Transactions on Intelligent Transportation Systems, 2023.

[88] J. Fu, W. Zhou, and J. Xu, "Design of improved PBFT algorithm based on aggregate signature and node reputation," Intelligent and Converged Networks, 2023.

[89] X. Fu, H. M. Wang, P. C. Shi, and X. H. Zhang, "Teegraph: A blockchain consensus algorithm based on TEE and DAG for data sharing in IoT," Journal of Systems Architecture, 2022.

[90] A. R. Harish, X. L. Liu, X. Wang, S. L. Pan, H. N. Dai, M. Li, and G. Q. Huang, "Blockchain FOR LOGISTICS 4.0: A SYSTEMATIC REVIEW AND PROSPECTS," Transportation Research Part E: Logistics and Transportation Review, 2025.

[91] R. Jabbar, E. Dhib, A. B. Said, M. Krichen, N. Fetais, E. Zaidan, and K. Barkaoui, "Blockchain technology for intelligent transportation systems: A systematic literature review," IEEE Access, 2022.

[92] X. D. Jia, X. M. Song, and M. Sohail, "Effective consensus-based distributed auction scheme for secure data sharing in internet of things," Symmetry, 2022.

[93] S. A. Khan, S. Balusamy, M. P. A. Saviour, S. Bojjawar, C. M. Chidambaranathan, and D. S. Sofia, "Integrating blockchain technology with artificial intelligence to create scalable, reliable, and open decision support systems," 2025 Global Conference in Emerging Technology, GINOTECH 2025, 2025.

[94] Z. Li, J. Li, F. Nie, B. Zhang, and J. Guo, "Optimization of blockchain consensus mechanism based on DPOS," Proceedings of SPIE - The International Society for Optical Engineering, 2023.

[95] A. Liu, Q. Zhang, S. W. Xu, H. M. Feng, X. B. Chen, and W. Liu, "Qbiot: A quantum blockchain framework for IoT with an improved proof-of-authority consensus algorithm and a public-key quantum signature," Computers, Materials and Continua, 2024.

[96] Y. Liu, J. Gao, Y. Lu, R. Cao, L. Yao, Y. Xia, and D. Han, "Lightweight blockchain-enabled secure data sharing in dynamic and resource-limited UAV networks," IEEE Network, 2024.

[97] Z. Ma, F. Richard Yu, X. Jiang, and A. Boukerche, "Trustworthy traffic information sharing secured via blockchain in VANETs," DIVANet 2020 - Proceedings of the 10th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications, 2020.

[98] Y. Meng, "Research on trusted exchange and data traceability mechanism of health information system integrating blockchain and internet of things," 2025 Asia Conference on Energy Conversion Systems and Power Electronics (AECSPE), 2025.

[99] S. S. F. Nasab, D. Bahrepour, and S. R. K. Tabbakh, "A review on secure data storage and data sharing technics in blockchain-based IoT healthcare systems," 2022 12th International Conference on Computer and Knowledge Engineering, ICCKE 2022, 2022.

[100] R. G. Reddy, P. S. Pateja, A. Patel, G. Palaniappan, and B. Rajendran, "Identifying sybil attacks in blockchain networks through behavioral analysis and zero knowledge proof implementations," 2024 1st International Conference on Data, Computation and Communication, ICDCC 2024, 2024.

[101] F. Ren and Z. Liang, "A data sharing privacy protection model based on federated learning and blockchain technology," International Journal of Advanced Computer Science and Applications, 2024.

[102] J. Ryu, Y. Lee, and Y. Yoon, "Blockchain model for reliable consensus algorithm on the autonomous driving data management," Digest of Technical Papers - IEEE International Conference on Consumer Electronics, 2024.

[103] A. Sarfaraz, R. K. Chakrabortty, and D. L. Essam, "The implications of blockchain-coordinated information sharing within a supply chain: A simulation study," Blockchain: Research and Applications, 2023.

[104] A. Shrivastava, R. Praveen, M. MuhsnHasan, S. Bansal, S. P. Dwivedi, and V, A, "Blockchain-powered secure data sharing in AI driven smart cities," 2025 International Conference on Computing and Communications (COMPUTINGCON), 2025.

[105] K. Singh, A. S. Rajawat, S. B. Goyal, and H. N. Waked, "A blockchain-based intelligent framework for secure and sustainable knowledge sharing in organizational systems," Procedia Computer Science, 2025.

[106] O. Ural and K. Yoshigoe, "Survey on blockchain-enhanced machine learning," IEEE Access, 2023.

[107] S. van Engelenburg, B. Rukanova, W. Hofman, J. Ubacht, Y. H. Tan, and M. Janssen, "Aligning stakeholder interests, governance requirements and blockchain design in business and government information sharing," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020.

[108] H. H. Wang, C. P. Wang, K. Zhou, D. Y. Liu, X. L. Zhang, and H. B. Cheng, "Tebchain: A trusted and efficient blockchain-based data sharing scheme in UAV-Assisted IoV for disaster rescue," IEEE Transactions on Network and Service Management, 2024.

[109] H. Wang, D. Mao, Z. Chen, H. Rao, and Z. Li, "Blockchain-based decentralized federated learning model," 2023 4th International Conference on Information Science, Parallel and Distributed Systems, ISPDS 2023, 2023.

[110] Q. Wang, Z. Wu, and Y. Lu, "A multi-layer secure sharing framework for aviation big data based on blockchain," Future Internet, 2025.

[111] W. Z. Wang, D. X. Tian, X. T. Duan, and J. S. Zhou, "Rpbft: A scalable consensus mechanism for large blockchain systems," Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 2024.

[112] Y. Wang and X. H. Zhao, "Ig-pgft: A secure and efficient intelligent grouping PBFT consensus algorithm for the industrial internet of things," Cluster Computing, 2025.

[113] Z. Wang, B. Wen, and Z. Luo, "Towards on blockchain data privacy protection with cryptography and software architecture approach," Communications in Computer and Information Science, 2020.

[114] F. Y. Xu, S. H. Hu, Y. Sun, X. X. Hu, J. Qi, Y. F. Sun, and Z. J. Dong, "Fdss: Flight data sharing scheme based on blockchain with dynamic, secure and efficient consensus algorithm," Computer Networks, 2025.

[115] Y. Yuman, S. B. Goyal, A. S. Rajawat, M. Kumar, A. Shankar, F. Alhayan, and S. Basheer, "A blockchain-based solution for enhancing the efficiency and security of healthcare knowledge sharing systems in the era of industry 4.0," Wireless Networks, 2025.

[116] S. F. Zhang, Y. Liu, and X. R. Chen, "Bit problem: Is there a trade-off in the performances of blockchain systems?," Communications in Computer and Information Science, 2020.

[117] X. F. Zhang, W. B. Xia, Q. M. Cui, X. F. Tao, and R. P. Liu, "Efficient and trusted data sharing in a sharding-enabled vehicular blockchain," IEEE Network, 2023.

[118] X. Zhang, Q. Guo, X. Pan, D. Bai, X. Pan, and J. Lv, "Study on the development status and prospect of blockchain technology in the energy field," ACM International Conference Proceeding Series, 2022.

[119] X. Zhang, M. Xue, and X. Miao, "A consensus algorithm based on risk assessment model for permissioned blockchain," Wireless Communications and Mobile Computing, 2022.

[120] Y. H. Zhang, Y. T. Wu, T. Li, H. Zhou, and Y. L. Chen, "Vertical federated learning based on consortium blockchain for data sharing in mobile edge computing," CMES - Computer Modeling in Engineering and Sciences, 2023.

[121] X. K. Zhou, W. Huang, W. Liang, Z. Yan, J. H. Ma, Y. Pan, and K. I. K. Wang, "Federated distillation and blockchain empowered secure knowledge sharing for internet of medical things," Information Sciences, 2024.

Generative AI in Higher Education: A Systematic Review with Emphasis on Programming and Computer Science Education

Abdulaziz Saidu Yalwa[1], Mohd Shahizan Othman[2], Lizawati Mi Yusuf[3], Muteb Sinhat Almarshadi[4]
Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor Bahru, Malaysia[1, 2, 3]
Department of Computer Sciences and Human Studies, Shaqra University, Shaqra, Kingdom of Saudi Arabia[4]

***Abstract*—Generative Artificial Intelligence (GenAI) is rapidly reshaping teaching and learning in higher education, particularly in programming and computer science education, where tools such as ChatGPT and GitHub Copilot are increasingly used for code generation, debugging, conceptual explanation, and personalised learning support. Despite this growing use, the literature remains fragmented, with limited synthesis across educational outcomes, adoption contexts, user perceptions, and factors influencing sustained use. To address this gap, this study conducts a systematic literature review following PRISMA 2020 guidelines. The review draws on a dataset of 60 empirical studies published between 2022 and 2025 and retrieved from major academic databases. The selected studies were analysed using a mixed synthesis approach that combines descriptive mapping with thematic analysis. The synthesis shows that GenAI can improve coding support, debugging efficiency, conceptual understanding, and student engagement, especially in programming-related learning contexts; however, adoption remains uneven across regions, institutions, and course settings, while concerns related to academic integrity, over-reliance, reliability, and unequal access persist. By integrating findings across learning outcomes, adoption patterns, user perceptions, and continuance-related factors, this review provides a more structured understanding of GenAI use in higher education, with particular emphasis on programming and computer science education. The study highlights the need for AI literacy, ethical governance, and inclusive institutional support to enable more responsible and sustainable GenAI integration.**

Keywords*—*Generative artificial intelligence; higher education; programming education; computer science education; continuance intention; technology acceptance model; expectation–confirmation model; PRISMA; systematic literature review

I. Introduction

Generative Artificial Intelligence (GenAI) has emerged as a transformative technology across multiple domains, including education, healthcare, finance, and software development[1]. In the context of higher education, particularly in programming and computer science education, GenAI tools such as ChatGPT and GitHub Copilot are increasingly used to support code generation, debugging, and real-time explanation of programming concepts[2], [3], [4], [5], [6], [7]. More broadly, higher education studies also report positive perceptions of usefulness, engagement, and instructional support [1], [8], [9]. These tools provide interactive and personalised learning experiences, which are especially beneficial for novice learners who often struggle with syntax, algorithmic thinking, and debugging processes.

The rapid adoption of GenAI has raised concerns regarding the reliability of AI-generated outputs, risks to academic integrity, and students' over-reliance on automated solutions [10]. Although several studies report improvements in coding accuracy, engagement, and motivation, issues such as incorrect code generation, misuse in assessments, and unequal access across institutions remain critical [3]. These challenges highlight that the integration of GenAI is not purely a technological issue but also a pedagogical and ethical concern.

Despite the growing body of research on GenAI in education, existing studies remain fragmented and predominantly focused on short-term learning outcomes. Limited attention has been given to the sustained use of these technologies in higher education, particularly in programming and computer science education. However, the long-term educational value of GenAI depends on whether students and educators continue to use these tools effectively over time. This concept, known as continuance intention, is explained through established models such as the Technology Acceptance Model (TAM), the Unified Theory of Acceptance and Use of Technology (UTAUT), and the Expectation–Confirmation Model (ECM), which link perceived usefulness, satisfaction, and behavioural intention to continued technology use [11].

Unlike previous reviews, this study integrates continuance intention theory with empirical evidence to provide a more comprehensive understanding of sustained GenAI use in higher education, particularly in programming and computer science education. By combining theoretical perspectives with observed educational outcomes, this study aims to bridge the gap between initial adoption and long-term utilisation of GenAI tools.

To address the lack of consolidated evidence, this study conducts a systematic literature review following PRISMA 2020 guidelines. The review synthesises findings from 60 empirical studies published between 2022 and 2025, focusing on learning outcomes, adoption contexts, user perceptions, and pedagogical challenges. Furthermore, the study develops a conceptual framework linking adoption conditions, usage mechanisms, educational outcomes, and associated challenges. By providing a structured and comprehensive synthesis, this research offers practical insights into the responsible and sustainable integration of GenAI in programming education.

While the review is anchored in programming and computer science education, it also draws on broader higher education studies that provide relevant empirical evidence on adoption,

governance, user perceptions, and sustained use conditions surrounding GenAI integration.

II. Research Background

Higher education is increasingly being shaped by digital technologies that support more interactive, flexible, and student-centered forms of learning. Within this broader context, programming and computer science education remain especially important because they aim to develop not only coding proficiency but also computational thinking, problem-solving, and algorithmic reasoning required for real-world application[12]. In response to evolving technological demands, teaching in these fields has progressively shifted toward active and experiential learning approaches that emphasise hands-on practice, timely feedback, and applied problem-solving.

Against this background, Generative Artificial Intelligence (GenAI) has emerged as a significant development in higher education, with particularly strong relevance for programming and computer science education. Tools such as ChatGPT and GitHub Copilot support code generation, debugging, conceptual explanation, and real-time assistance during learning tasks [4], [13], [14]. These capabilities make learning more adaptive and personalised, especially for novice programmers who often benefit from continuous guidance, immediate feedback, and opportunities to explore alternative coding approaches. As a result, GenAI is increasingly being viewed as a flexible instructional support tool that can enhance coding efficiency, conceptual understanding, collaboration, and learner engagement while complementing more traditional teaching methods [1], [8].

However, the integration of GenAI also introduces important pedagogical, ethical, and institutional challenges. Concerns have been raised about over-reliance on AI-generated solutions, inaccuracies in generated outputs, and issues related to academic integrity and fairness [15], [16]. In addition, differences in infrastructure, access, and institutional readiness contribute to uneven adoption across educational settings and regions [17]. These issues show that GenAI is not only a technical development but also one that must be understood through pedagogical, ethical, and institutional perspectives.

Beyond initial adoption, the continued use of GenAI has become an important issue in higher education, particularly in programming and computer science education [18]. Continuance intention provides a useful lens for understanding whether students and lecturers remain willing to use GenAI tools over time. In this regard, the Expectation–Confirmation Model (ECM) highlights the importance of satisfaction and perceived usefulness in shaping continued engagement [11]. Although this perspective is increasingly relevant, much of the existing literature still focuses on short-term outcomes rather than the conditions that support sustained and responsible use.

Despite the growing number of studies on GenAI in higher education, particularly in programming and computer science education, the literature remains fragmented and lacks a sufficiently integrated perspective. Many studies focus on immediate educational outcomes, while giving less attention to sustained use, contextual variation, and the interaction of technological, pedagogical, and ethical factors. This study addresses these gaps by providing a systematic synthesis of empirical evidence, integrating continuance-related perspectives, and developing a more connected understanding of how GenAI is being adopted, experienced, and sustained in educational contexts.

A. Challenges and Ethical Concerns of GenAI Integration

Even with its potential, the integration of GenAI introduces several challenges. One major concern is over-reliance on AI-generated solutions, which may limit the development of independent problem-solving skills among students[12]. In addition, inaccuracies in generated outputs can mislead learners, particularly those with limited programming experience.

Academic integrity is another critical issue, as AI-generated code can be difficult to distinguish from students' original work, raising concerns about assessment validity and fairness. Furthermore, disparities in infrastructure, access to advanced tools, and institutional readiness contribute to uneven adoption across regions and educational contexts[13].

These challenges highlight that the integration of GenAI is not purely a technical issue but also involves ethical, pedagogical, and institutional dimensions that must be addressed to ensure responsible use.

B. Continuance Intention and Sustained Educational Use of GenAI

Beyond initial adoption by [14], sustained use of GenAI remains a critical issue in higher education, particularly in programming and computer science education. Continuance intention explains users' willingness to continue using a technology after initial adoption and is commonly modelled using the Expectation–Confirmation Model (ECM) [16]. According to ECM, continued use is influenced by satisfaction, perceived usefulness, and confirmation of expectations.

While models such as the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) explain initial adoption behaviour [19] ECM provides a stronger theoretical basis for understanding post-adoption continuance. In the context of GenAI, continuance intention reflects the extent to which students and lecturers are willing to repeatedly use tools such as ChatGPT and GitHub Copilot for learning and instructional purposes.

Empirical evidence suggests that factors such as perceived usefulness, trust, satisfaction, and ethical awareness play important roles in shaping sustained engagement. Therefore, continuance intention provides a critical lens for analysing the long-term integration of GenAI in programming education.

C. Benefits and Educational Capabilities of GenAI

GenAI tools provide scalable support for programming education through automated code generation, debugging assistance, and real-time explanation of programming concepts[8]. By analysing user inputs and coding patterns, these tools deliver adaptive feedback and personalised learning support that enhances engagement, coding efficiency, and conceptual understanding [1].

In addition, GenAI supports exploration and collaborative learning by allowing students to experiment with different coding approaches and interact with intelligent systems that simulate real-world development practices. These capabilities

position GenAI as a flexible and learner-centered support system that complements traditional teaching methods.

D. Summary of Research Gaps

Despite the increasing number of studies on GenAI in higher education, particularly in programming and computer science education, the literature remains fragmented and lacks a unified perspective. Existing research primarily focuses on short-term learning outcomes, with limited attention to sustained use and long-term educational impact.

Moreover, theoretical integration remains insufficient, as many studies do not explicitly connect empirical findings with established technology adoption models such as ECM, TAM, and UTAUT. In addition, challenges related to ethics, pedagogy, and infrastructure are often discussed in isolation rather than as interconnected factors.

This study addresses these gaps by providing a systematic synthesis of empirical evidence, integrating continuance intention theory, and developing a conceptual framework that links adoption contexts, usage mechanisms, outcomes, and challenges for sustainable GenAI integration in higher education, particularly in programming education.

III. Methodology

This study adopts a systematic literature review (SLR) approach to synthesise empirical evidence on the use of Generative Artificial Intelligence (GenAI) in higher education, with particular emphasis on programming and computer science education. The review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure methodological transparency, reproducibility, and systematic reporting of the review process [20].

The overall procedure consists of the following structured stages:

- Definition of research objectives and research questions
- Development of a review protocol, including search strategy and selection criteria.
- Systematic literature search across multiple academic databases.
- Screening and selection of relevant studies based on predefined criteria.
- Data extraction using a structured framework.
- Quality assessment of selected studies; and
- Thematic synthesis and analysis of findings.

This structured methodology ensures a transparent and replicable review process suitable for high-quality academic research.

A. Research Questions

To structure the analysis and ensure a focused investigation, this systematic literature review is guided by the following research questions.

- RQ1. How do GenAI tools influence learning outcomes in higher education, particularly in programming and computer science education, including code correctness, debugging efficiency, conceptual understanding, and student engagement?
- RQ2. What differences exist in GenAI adoption across countries, institutions, course levels, and access conditions such as infrastructure, policies, and licensing availability?
- RQ3. What are the perceptions of students and lecturers toward GenAI in higher education, particularly in programming and computer science education, and which factors influence their continued intention to use these technologies?
- RQ4. What pedagogical, ethical, and academic integrity challenges arise from integrating GenAI into higher education, particularly in programming and computer science education, and what mitigation strategies have been proposed?

B. Review Protocol Design

To ensure the review is conducted in a systematic and reliable way, a review protocol was prepared before starting the data collection. This protocol served as a guide for the entire review process and helped to reduce possible bias.

It clearly defined the search strategy, including the keywords and Boolean combinations used to retrieve relevant studies. It also specified the selection of appropriate academic databases and established the inclusion and exclusion criteria to determine which studies should be considered.

In addition, the protocol outlined how data would be extracted from the selected studies in a consistent manner. A quality assessment approach was also included to evaluate the strength and relevance of each study. Furthermore, the protocol described how the collected data would be analysed and organised to ensure a clear and structured synthesis of the findings.

Overall, this protocol improves the consistency, transparency, and reliability of the review, making it easier for the process to be understood and replicated.

C. Data Sources and Search Strategy

A comprehensive literature search was conducted across Scopus, Web of Science, IEEE Xplore, and SpringerLink. These databases were selected because they index high-quality peer-reviewed research across computer science, education, and educational technology disciplines. The search included studies published between January 2022 and September 2025, reflecting the recent emergence and rapid development of GenAI in higher education.

The following Boolean search string was applied across the databases, with minor formatting adjustments where required by database syntax: ("Generative Artificial Intelligence" OR "GenAI" OR "ChatGPT" OR "GitHub Copilot") AND ("higher education" OR "university" OR "college") AND ("programming" OR "computer science" OR "coding").

This search strategy was designed to prioritise empirical studies on the use, adoption, and educational impact of GenAI in programming and computer science education within higher education. Relevant higher education studies outside strictly programming-specific contexts were also retained during screening where they provided direct insight into adoption conditions, user perceptions, policy environments, ethical concerns, or sustained use factors relevant to GenAI integration.

D. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected studies, clear inclusion and exclusion criteria were defined before the screening process. Only studies that reported empirical findings on the use of Generative Artificial Intelligence (GenAI) in programming or computer science education were considered. The review focused specifically on studies involving higher education students or lecturers, as this aligns with the scope of the study.

Although this review primarily focuses on programming and computer science education, a small number of broader higher education studies were included when they provided empirical evidence directly relevant to GenAI adoption, user perceptions, governance, institutional policies, ethical considerations, or sustained use. These studies were retained because they offer insights into contextual and continuance-related factors that inform the integration of GenAI in programming-focused learning environments.

In addition, only articles published in peer-reviewed journals or conference proceedings between 2022 and 2025 were included, reflecting the recent and rapidly evolving nature of GenAI in education. To maintain consistency and clarity in analysis, only studies published in English were considered.

On the other hand, studies were excluded if they did not meet these criteria. This includes publications such as editorials, opinion papers, conceptual papers, and review articles, as they do not provide primary empirical evidence. Studies published outside the selected time frame or those not directly related to GenAI use in higher education, particularly in programming and computer science education, were also excluded.

These criteria helped to narrow the literature to relevant and high-quality studies, ensuring that the review is focused, consistent, and academically robust.

E. Screening Process

The database search initially identified 3,160 records. After removing 956 duplicates, 2,204 unique records remained for screening. Title and abstract screening excluded 2,100 records that did not meet the study criteria. The remaining 104 articles were assessed through full-text review, resulting in 60 studies that satisfied the inclusion criteria and were retained for the final synthesis. The complete screening process is presented in the PRISMA flow diagram Fig. 1.

F. Data Extraction and Review Reliability

To ensure consistency and transparency in the review process, a structured data extraction form was developed for all selected studies. This form was designed to systematically capture relevant information from each article in a clear and organized manner.

The extracted data included key bibliographic information, as well as important analytical details such as the study context, research design, sample characteristics, and the types of GenAI tools used. In addition, information related to learning outcomes, user perceptions, reported challenges, and proposed mitigation strategies was collected to support a comprehensive analysis.

To improve the reliability of the data extraction process, two researchers independently reviewed the extracted information. Any differences in interpretation or classification were discussed and resolved collaboratively. This approach helped to reduce bias and ensure consistency in the analysis[21].

Furthermore, the reporting of the review process was guided by the PRISMA 2020 framework, which supports transparency and allows the review procedure to be clearly understood and replicated.

G. Quality Assessment

To ensure the credibility and methodological rigor of the selected studies, a structured quality assessment was carried out using a checklist adapted from the Critical Appraisal Skills Programme (CASP), which is commonly used in systematic reviews across interdisciplinary research[5].

The assessment focused on five key aspects: the clarity of the research objectives, relevance to the study focus, methodological rigor, contribution to theory or practice, and transparency in reporting limitations. Each study was evaluated using a binary scoring system (0–1) for each criterion, resulting in a maximum possible score of 5. To maintain a consistent quality standard, only studies with a minimum score of 3 were included in the final analysis.

To improve reliability and reduce subjective bias, the quality assessment was conducted independently by two reviewers. Any differences in scoring were discussed and resolved collaboratively, and where necessary, a third reviewer was consulted to reach a final decision. This process helped to ensure consistency and strengthen the reliability of the evaluation.

The quality assessment played an important role in filtering out studies with unclear objectives, weak methodology, or limited relevance to GenAI in higher education, particularly in programming and computer science education. As a result, the final set of studies included in this review consists of high-quality research that provides meaningful insights into the adoption, impact, sustained use, and pedagogical implications of Generative Artificial Intelligence in higher education, particularly in programming and computer science education.

H. Coding and Thematic Analysis Approach

A mixed synthesis approach, combining both qualitative and quantitative content analysis, was employed to analyse the selected studies, allowing a comprehensive examination of both descriptive trends and deeper thematic patterns within the literature. The coding process was conducted in two stages to ensure reliability and transparency. First, descriptive coding classified each study according to publication year, country or region, publication source, course level, GenAI tool, research design, and participant group. Second, analytical coding identified themes directly aligned with the research questions, including learning outcomes, debugging support, conceptual understanding,

adoption conditions, user perceptions, continuance intention, academic integrity, ethical concerns, and mitigation strategies. For example, studies reporting improvements in code accuracy or debugging efficiency were coded under "learning outcomes," while studies discussing trust, satisfaction, or continued use were coded under "continuance intention."

To ensure consistency, a random subset of 20 articles was independently coded by two reviewers, achieving an inter-rater agreement of over 94%. Any disagreements in coding were discussed until consensus was reached. This dual coding strategy enabled the review to capture both the breadth of adoption patterns and the depth of thematic insights, providing a well-rounded understanding of the role of GenAI in higher education, particularly in programming and computer science education.

IV. Findings

This section presents the thematic findings of the review based on the synthesis of 60 empirical studies examining the use of Generative Artificial Intelligence (GenAI) in higher education, with particular emphasis on programming and computer science education. The reviewed studies comprised quantitative, qualitative, and mixed-method designs, enabling a comprehensive synthesis of adoption patterns, user perceptions, and educational outcomes associated with GenAI use in higher education, particularly in programming contexts.

A. Impact on Learning Outcomes and Engagement

The reviewed studies consistently indicate that GenAI tools are increasingly used in higher education, particularly in programming and computer science education, to support code generation, debugging, and personalised learning assistance. Most studies indicate that these tools contribute positively to student learning outcomes by improving coding accuracy, accelerating debugging processes, and increasing learners' confidence in programming tasks [4]. In addition, GenAI-supported feedback systems provide real-time guidance that helps reduce frustration and sustain student engagement when addressing complex programming concepts [22], [23]. Adaptive features further enable students to progress at their own pace while addressing individual knowledge gaps [6].

Beyond individual performance gains, studies highlight the role of GenAI in supporting collaborative and reflective learning. The use of AI tools in group programming activities has been shown to stimulate peer discussion, collaborative debugging, and critical reflection on AI-generated outputs, thereby strengthening both cognitive and social aspects of programming education [10], [24]. In addition, personalised feedback mechanisms may help reduce performance disparities between students with different levels of programming proficiency [6].

However, the effectiveness of GenAI tools varies across platforms, programming tasks, and instructional contexts. Comparative studies show that the accuracy and usefulness of tools depend on the complexity of the coding problem and the level of student expertise [2]. Many studies report improvements in motivation and engagement; evidence regarding consistent improvements in academic performance remains mixed [8]. Some research also indicates that GenAI tools may struggle with complex coding problems, highlighting the need for educators to critically guide students in evaluating and adapting AI-generated solutions [25].

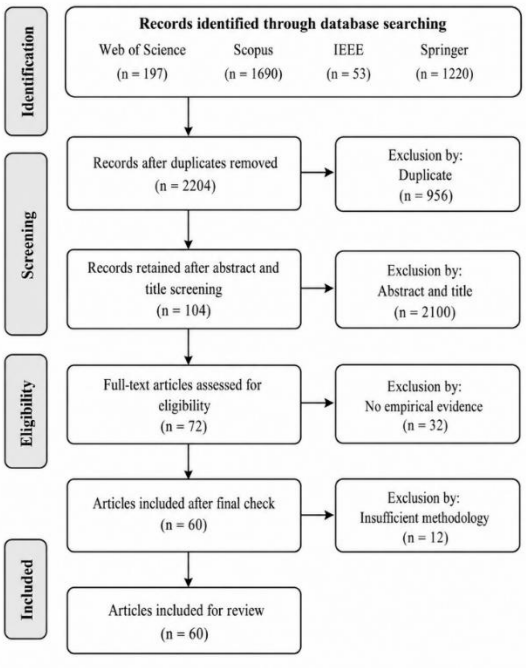


Fig. 1. PRISMA 2020 flow diagram illustrating the study selection process.

The surge in publications coincides with the emergence of large language model–based systems such as ChatGPT and GitHub Copilot, which have significantly accelerated experimentation with AI-assisted programming learning. No empirical studies were identified in 2022, reflecting the late emergence of GenAI research following the release of ChatGPT in late 2022. The zero count for 2022 reflects the fact that mainstream pedagogical research into generative AI was catalyzed by the release of ChatGPT in November 2022, meaning that the first wave of empirical findings did not reach publication until 2023. Followed by 6 in 2023 and 22 in 2024, reflecting a substantial increase in scholarly attention. By 2025, an additional 32 studies had already been reported, suggesting that research on GenAI in higher education, particularly in programming and computer science education, continues to expand rapidly, as shown in Fig. 2.

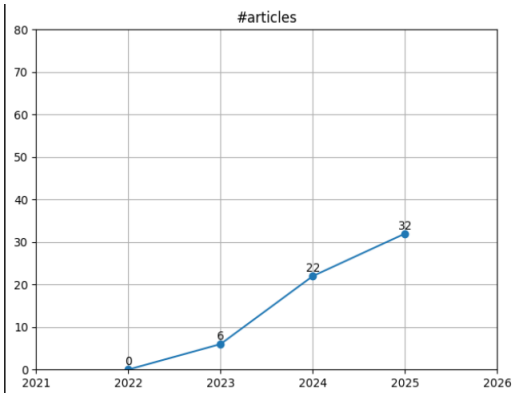


Fig. 2. Year-wise distribution of publications on GenAI in programming education (2022-2025).

B. Contexts of Adoption and Equity

The reviewed studies reveal notable disparities in the adoption of Generative Artificial Intelligence (GenAI) across course levels, institutions, and geographical contexts. Although many studies examine GenAI use in higher education broadly, fewer focus on specific domains within computer science education.

Introductory programming (CS1) emerges as the most common context of application, with 12 studies (20%) reporting the use of GenAI tools to support beginners in learning fundamental coding skills. Advanced programming and software engineering courses account for 10 studies (16.6%), while algorithms and data structures are addressed in 8 studies (13.3%). In addition, 18 studies (30%) investigate GenAI adoption in higher education settings without specifying a particular programming domain.

These patterns suggest that GenAI tools are currently used most extensively in introductory programming environments, where real-time guidance can assist students in addressing syntax errors, debugging tasks, and foundational programming concepts. The distribution of studies across computer science domains is illustrated in Fig. 3, highlighting the concentration of GenAI applications at early stages of programming education.

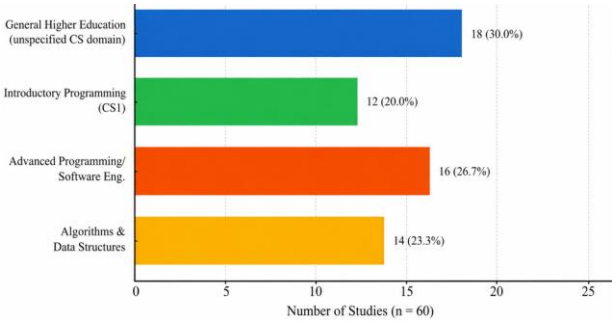


Fig. 3. Distribution of reviewed studies across computer science domains.

Regional disparities were also evident across the reviewed studies. Higher levels of GenAI adoption were reported in countries such as China, the United States, and Australia, where strong institutional support, technological infrastructure, and access to AI tools have enabled early integration into programming education [1], [39], [9]. In contrast, studies from Africa and South Asia highlight barriers related to limited infrastructure, policy readiness, and access to AI resources, which constrain the equitable adoption of GenAI in higher education [26], [27]. These findings suggest that institutional capacity and resource availability play an important role in shaping the global diffusion of GenAI in programming education. As illustrated in Fig. 4, China accounts for the largest share of studies (12 studies, approximately 20%), indicating its leading position in research and implementation within this domain.

The findings indicate that although the adoption of GenAI in programming education is expanding, it remains concentrated in introductory course levels and technologically advanced regions. This uneven distribution highlights potential disparities in access and implementation across institutions and regions. If resource-limited contexts are not adequately supported, the diffusion of GenAI may further widen existing global inequalities in computing education. Ensuring equitable access to technological infrastructure, policy support, and digital literacy therefore remains essential for the sustainable and inclusive integration of GenAI in programming education.

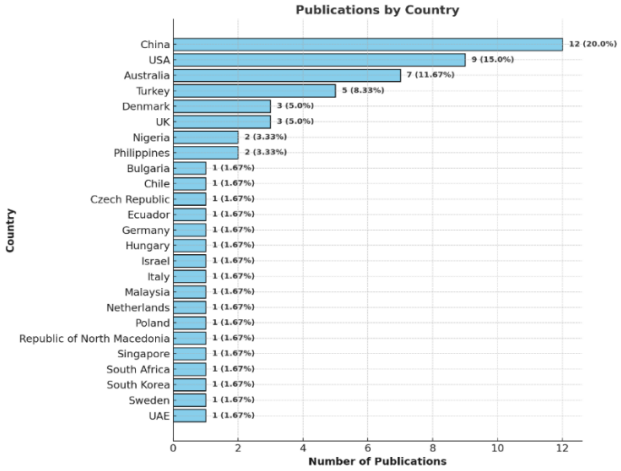


Fig. 4. Geographic distribution of reviewed studies by country.

C. Students' and Lecturers' GenAI Perceptions and Continuance of Use

Most reviewed studies report that students hold generally positive perceptions of GenAI tools, describing them as supportive resources that reduce frustration and make programming tasks more engaging see Fig. 5. Students frequently highlight the usefulness of tools such as ChatGPT and GitHub Copilot in clarifying difficult programming concepts, assisting with debugging, and strengthening confidence in coding abilities [26]. These findings suggest that learners perceive GenAI not merely as a convenience but as a learning aid that supports persistence and self-efficacy in programming education.

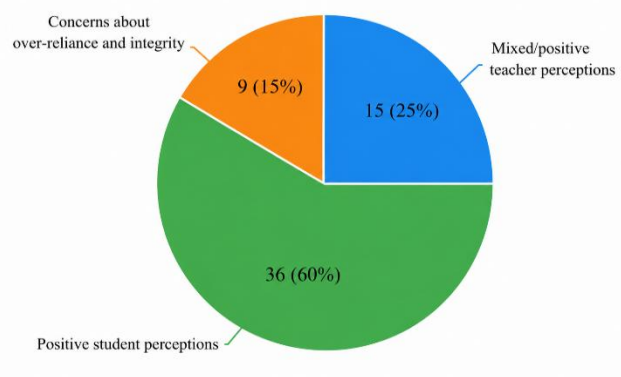


Fig. 5. Comparison of students' and lecturers' perceptions of GenAI use.

Lecturers, however, express more cautious and divided perspectives. Some acknowledge that GenAI can improve instructional efficiency by assisting with automated feedback and assessment processes [1]. Others raise concerns related to academic integrity, fairness in assessment, and the potential for students to become overly dependent on AI-generated solutions [28]. Questions about the reliability of AI-generated code further contribute to this caution, as inaccuracies or hidden errors may affect trust in GenAI for complex or high-stakes programming tasks [2].

These findings reveal a perceptual gap between students and lecturers. While students tend to view GenAI as an empowering learning tool and show strong willingness to continue using it, lecturers often weigh its efficiency benefits against concerns related to trust, originality, and long-term learning outcomes. This divergence suggests that continuance intention is influenced not only by perceived usefulness but also by factors such as trust, ethical considerations, and the pedagogical frameworks established by educators.

D. Pedagogical and Integrity Challenges

The review identified several challenges that complicate the integration of GenAI into programming education, particularly regarding tool reliability and academic integrity. A recurring concern is the accuracy of AI-generated code. Although tools such as ChatGPT can quickly generate functional code snippets, studies report that the outputs may contain logical or syntactic errors that novice learners often struggle to detect [2]. Such limitations can mislead students and highlight the need for critical evaluation of AI-generated solutions. In addition, excessive reliance on GenAI tools may reduce opportunities for students to develop independent debugging and problem-solving skills, raising questions about their long-term pedagogical impact [12].

Academic integrity represents another major challenge. Several studies indicate that GenAI tools can be misused to bypass programming assessments, blurring the boundary between authentic student work and AI-assisted outputs [29]. This creates concerns about fairness, transparency, and the validity of assessment when AI-generated code becomes difficult to distinguish from student submissions [59]. These issues highlight the importance of establishing clear institutional policies and pedagogical guidelines to ensure responsible use of GenAI in programming courses.

Finally, the findings indicate that the integration of GenAI in higher education, particularly in programming and computer science education, is expanding rapidly but remains uneven across institutional contexts and geographical regions. While many studies highlight improvements in coding support, debugging efficiency, and learner engagement, concerns regarding academic integrity, over-reliance on AI-generated solutions, and unequal access to technological infrastructure continue to shape how these tools are adopted in educational settings. These patterns underscore the need for pedagogically guided and ethically responsible integration of GenAI within programming curricula.

V. Discussion

This section synthesises the key findings of the review, explicitly relates them to the research questions (RQ1–RQ4), and discusses their broader implications for theory, practice, and policy in higher education, with particular emphasis on programming and computer science education. It further identifies critical gaps and outlines directions for the responsible and sustainable use of GenAI in higher education.

A. Influence of GenAI on Learning Outcomes in Higher Education

In response to RQ1, which examines how GenAI tools influence programming learning outcomes, the findings consistently indicate that GenAI significantly improves code correctness, debugging efficiency, conceptual understanding, and student engagement. This aligns with prior studies [6], [2], which report that AI-assisted programming tools enhance students' coding accuracy and reduce cognitive load during problem-solving tasks.

GenAI tools function as intelligent learning assistants, providing real-time feedback, automated code suggestions, and explanatory support that facilitate students' problem-solving processes [21]. These capabilities contribute to increased confidence and engagement, reinforcing both cognitive and affective dimensions of learning. Similar findings have been reported in [8], where students using GenAI demonstrated higher levels of motivation and deeper conceptual understanding compared to traditional learning environments.

However, the findings also reveal that these benefits are context-dependent. While performance improvements are evident, excessive reliance on AI-generated solutions may reduce active learning and limit the development of independent problem-solving skills. This concern is consistent with studies by [8] and [24], which highlight the risk of overdependence on AI tools leading to superficial learning and reduced critical thinking abilities.

This suggests that the impact of GenAI is not inherently positive but is mediated by how it is integrated into pedagogical practices. As noted in [23], effective instructional design and guided use of GenAI are essential to ensure that students engage meaningfully with learning tasks rather than passively relying on automated outputs.

B. Variability in GenAI Adoption Across Contexts

Addressing RQ2, the review reveals substantial variability in the adoption of GenAI across countries, institutions, course levels, and access conditions, including digital infrastructure, institutional policies, and licensing availability. This finding is consistent with prior studies [59], [3], which emphasize that the uptake of emerging educational technologies is highly contingent on contextual and organizational factors.

The findings suggest that GenAI adoption is uneven and shaped by a combination of technological readiness and institutional support. Institutions with well-developed digital infrastructure, stable internet access, and clearly defined policy frameworks tend to demonstrate higher levels of GenAI integration in teaching and learning practices. In contrast, resource-constrained environments face persistent barriers related to limited access, high subscription costs, and insufficient technical support. Similar patterns have been reported in [11], where disparities in infrastructure and funding significantly influenced the adoption of AI-based educational tools.

This variability reflects broader issues of digital inequality, where unequal access to GenAI tools risks reinforcing existing educational disparities between institutions and regions. As highlighted in [59], students and educators in under-resourced settings may be systematically disadvantaged, limiting their ability to benefit from AI-enhanced learning environments.

Furthermore, inconsistent or ambiguous institutional policies create uncertainty regarding acceptable use, academic integrity, and ethical boundaries, which may hinder effective and

confident adoption. This aligns with findings from [17] and [8], which indicate that the absence of clear governance frameworks often leads to cautious or restricted use of GenAI by educators.

Taken together, these findings suggest that GenAI adoption should not be viewed as a purely technological issue, but rather as a context-dependent and policy-sensitive process. Effective integration, therefore, requires not only access to technology, but also institutional readiness, clear guidelines, and equitable resource distribution, as emphasized in [17].

C. Perceptions and Continuance Intention of GenAI Use

The findings indicate that both students and lecturers perceive GenAI as a useful and efficient tool that enhances learning, problem-solving, and conceptual understanding [12], [30]. However, continuance intention is not driven by perceived usefulness alone but by interrelated factors including satisfaction, trust, motivation, and ethical alignment.

Consistent with the Expectation-Confirmation Model (ECM), satisfaction emerges as a key determinant of continued use, shaped by post-adoption experiences [31], [11]. Trust also plays a critical role, particularly when students are perceived to use GenAI responsibly and in line with academic integrity standards [18], [32].

Ethical considerations further reinforce this process, as clear guidelines and training enhance both academic integrity and trust, supporting sustainable adoption [10]. Additionally, GenAI positively influences student motivation and engagement, indirectly strengthening continuance intention [33].

The findings suggest that continuance intention differs between students and lecturers. For students, continued use is mainly shaped by perceived usefulness [30], immediate learning support, confidence in programming tasks, debugging assistance, and reduced frustration. In contrast, lecturers' continuance intention is more strongly influenced by trust, reliability of AI-generated outputs, academic integrity, assessment fairness, institutional policy, and the extent to which GenAI supports rather than weakens learning objectives. Therefore, sustainable GenAI adoption requires addressing both student learning benefits and lecturer concerns regarding responsible use.

D. Pedagogical, Ethical, and Academic Integrity Challenges

In response to RQ4, the review identifies key challenges associated with GenAI integration in programming education, including academic integrity risks, over-reliance, reliability concerns, and ethical misuse. These challenges highlight that GenAI adoption extends beyond technical benefits to important pedagogical and ethical considerations.

Academic integrity is a major concern, as GenAI can generate complete solutions, making it difficult to assess students' original work. This challenges traditional assessment practices and raises issues of fairness and authenticity, as also reported in [34].

Over-reliance on GenAI may limit the development of essential programming skills, such as problem-solving and algorithmic thinking. While GenAI improves efficiency, excessive dependence can reduce active learning, consistent with findings in [21], [6].

Reliability issues further complicate its use, as GenAI may produce incorrect or misleading outputs. Without proper verification, this can negatively affect learning outcomes, particularly for less experienced students [6].

Ethical misuse is another critical issue, as unclear guidelines may lead to inappropriate use, including plagiarism or over-dependence on AI-generated content. This underscores the importance of ethical awareness and governance in GenAI adoption [33].

Although GenAI tools are useful for introductory programming tasks, their effectiveness is more limited in advanced programming contexts. Complex algorithmic reasoning, software architecture design, large-scale debugging, security-sensitive coding, and domain-specific optimisation tasks often require deeper human judgement. GenAI-generated code may appear syntactically correct while containing hidden logical errors, inefficient structures, or unsuitable design assumptions.

This review is also subject to temporal limitations. GenAI tools and their educational applications are evolving rapidly, and findings from earlier studies, particularly those published in 2022 and 2023, may not fully reflect the current capabilities of newer GenAI systems or recent institutional policy developments. As a result, the conclusions of this review should be interpreted as a synthesis of the available evidence within the 2022–2025 publication window rather than a final or static account of GenAI use in higher education. Therefore, GenAI should be used as a guided learning assistant rather than a replacement for independent reasoning, expert feedback, and rigorous code verification.

To address these challenges, the literature suggests redesigning assessments to focus on learning processes, promoting AI literacy, integrating GenAI within guided learning environments, and establishing clear institutional policies.

Overall, these findings indicate that addressing ethical and academic integrity concerns is essential for fostering trust and ensuring the sustainable and responsible use of GenAI in higher education.

E. Implications for Theory, Practice, and Policy

The findings of this study offer important implications across theoretical, practical, and policy dimensions. From a theoretical perspective, the results highlight the need to extend existing technology adoption models by integrating ethical and pedagogical factors, particularly in explaining continuance intention in GenAI-supported learning environments. This supports the inclusion of constructs such as trust, academic integrity, and ethical considerations beyond traditional frameworks like TAM and UTAUT [19].

From a practical perspective, educators are encouraged to adopt balanced instructional strategies that leverage the benefits of GenAI while preserving active learning, critical thinking, and independent problem-solving skills. This aligns with prior studies emphasizing guided and responsible AI integration in education [33]. Institutions should also provide adequate training, technical support, and clear usage guidelines to ensure effective and responsible student engagement with GenAI tools.

From a policy perspective, the findings underscore the need for well-defined institutional and regulatory frameworks that address academic integrity, ethical use, and data governance in AI-assisted learning environments. Clear policies are essential to reduce ambiguity, enhance trust, and support sustainable adoption of GenAI in higher education [26], [27].

F. Research Gaps and Future Directions

The review reveals several important gaps in the current literature on Generative AI in education. First, longitudinal evidence remains limited, as most studies examine immediate or short-term outcomes rather than the sustained effects of GenAI on learning performance, skill development, and academic behaviour over time. As a result, the durability of the reported benefits is still unclear. Second, contextual variation across regions, institutions, and resource environments remains insufficiently addressed. Differences in infrastructure, policy support, institutional readiness, and access conditions are likely to shape both adoption and educational effectiveness, yet these factors are rarely examined in a systematic way.

This unevenness is also reflected in the distribution of studies presented in Table I. Although the reviewed literature spans multiple countries, the evidence remains concentrated in a relatively small number of national settings, with China appearing most frequently. This pattern likely reflects stronger investment in artificial intelligence, rapid digitalisation in higher education, and a large STEM-oriented academic community capable of supporting early experimentation and sustained publication. A similar concentration is evident at the technological level, where ChatGPT overwhelmingly dominates the reviewed studies. Its accessibility, ease of use, and broad applicability to tasks such as writing, coding, explanation, and problem-solving have made it the most common platform for empirical investigation. By contrast, other tools, including Copilot, Gemini, DALL-E, and Midjourney, appear only sporadically, suggesting that the current knowledge base is still being developed around a narrow set of technological experiences. Taken together, these patterns indicate that existing findings, while valuable, may reflect the conditions of well-resourced contexts and the affordances of a single dominant platform more than the full diversity of educational realities. Table I therefore not only describes the spread of literature but also underscores the need for broader comparative work across countries, institutional settings, and GenAI tools to strengthen the robustness and transferability of future evidence.

TABLE I. SUMMARY OF THE DISTRIBUTION OF THE STUDY

SN	*Author(s)*	*Ref.*	*Year*	*Country*	*Tools*
1	Ivanov et al.	[35]	2024	Bulgaria	ChatGPT
2	C. K. Y. Chan & Tsi	[28]	2024	China	ChatGPT
3	Yang et al.	[36]	2025	China	ChatGPT
4	Suchanek & Kralova	[31]	2025	Czech Republic	ChatGPT
5	Tbaishat et al.	[37]	2025	UAE	ChatGPT, Copilot
6	Yilmaz & Karaoglan Yilmaz	[5]	2023	Turkey	ChatGPT
7	Kohnke et al.	[38]	2023	China	ChatGPT
8	Chiu	[39]	2024	China	ChatGPT, Copilot
9	Lee et al.	[1]	2024	Australia	ChatGPT
10	Awidi	[23]	2024	Australia	ChatGPT
11	Kohnke	[40]	2024	China	ChatGPT
12	Ulla et al.	[41]	2024	Philippines	ChatGPT
13	Wang et al.	[9]	2024	USA	ChatGPT
14	Kong et al.	[34]	2024	China	ChatGPT, Sora
15	Jin et al.	[17]	2025	Australia	Midjourney
16	Huang et al.	[42]	2025	USA	ChatGPT
17	Hashmi & Bal	[43]	2024	USA	ChatGPT, DALL-E
18	Charles et al.	[44]	2025	Australia	ChatGPT, Copilot
19	Ukwandu et al.	[45]	2024	UK	ChatGPT
20	Zou et al.	[46]	2025	China	GenAI chatbot
21	Otto et al.	[25]	2025	Denmark	ChatGPT, Copilot
22	N. N. Chan et al.	[47]	2025	Malaysia	ChatGPT, Copilot
23	Gmiterek & Kotuła	[48]	2025	Poland	ChatGPT, Gemini
24	Beckman et al.	[49]	2025	Australia	ChatGPT
25	Zhao et al.	[50]	2024	China	ChatGPT
26	Ratten & Jones	[51]	2023	Australia	ChatGPT
27	Ostick et al.	[52]	2025	USA	ChatGPT, Gemini
28	Dai et al.	[53]	2023	China	ChatGPT
29	Cha et al.	[54]	2024	China	ChatGPT
30	Cho & Ofosu-Anim	[19]	2025	South Korea	ChatGPT
31	De Putter-Smits et al.	[55]	2025	Netherlands	ChatGPT
32	McDonald et al.	[56]	2025	USA	ChatGPT
33	V. C. Chan	[57]	2025	USA	ChatGPT
34	Nikolic et al.	[29]	2024	Australia	ChatGPT, Copilot
35	Bikanga Ada	[12]	2024	UK	ChatGPT
36	Llerena-Izquierdo et al.	[3]	2024	Ecuador	Gemini
37	Mellado et al.	[4]	2024	Chile	GenAI
38	Sarsam et al.	[58]	2023	UK	Bard GAI chatbot
39	Hernandez et al.	[6]	2025	Philippines	ChatGPT
40	Huynh & Aichner	[15]	2025	Italy	ChatGPT, DALL-E
41	Yakubu et al.	[26]	2025	Nigeria	ChatGPT, Copilot
42	Otto, Lavi, et al.	[59]	2025	Denmark	ChatGPT
43	Nicolajsen et al.	[60]	2024	Denmark	ChatGPT
44	Aruleba et al.	[10]	2025	South Africa	ChatGPT
45	Ren & Wu	[33]	2025	USA	ChatGPT
46	Cengiz & Peker	[61]	2025	Turkey	ChatGPT
47	Qu et al.	[62]	2024	Singapore	ChatGPT
48	Heil et al.	[63]	2025	Germany	ChatGPT

49	Kohen-Vacs et al.	[2]	2025	Israel	ChatGPT, Copilot
50	Bjelobaba et al.	[13]	2025	Sweden	ChatGPT
51	C. K. Y. Chan & Hu	[8]	2023	China	ChatGPT
52	Yusuf et al.	[27]	2024	Nigeria	ChatGPT
53	An et al.	[64]	2025	USA	ChatGPT
54	Zdravkova & Ilijoski	[65]	2025	Republic of North Macedonia	ChatGPT
55	Ursavaş et al.	[66]	2025	Turkey	ChatGPT, Copilot
56	Fan et al.	[67]	2025	China	ChatGPT
57	Alagöz Hamzaj	[22]	2025	Turkey	ChatGPT
58	Dabis & Csáki	[68]	2024	Hungary	ChatGPT
59	Akçapınar & Sidan	[7]	2024	Turkey	ChatGPT
60	Irish et al.	[24]	2025	USA	ChatGPT, Gemini

Third, existing studies tend to examine technological, pedagogical, and ethical dimensions in isolation, resulting in fragmented insights. This highlights the need for more integrated frameworks that capture the complex interactions between these factors.

Future research should therefore adopt interdisciplinary and mixed-method approaches to provide a more holistic understanding of GenAI use in education. In particular, studies should focus on developing unified models that incorporate technological, pedagogical, and ethical constructs to better explain sustainable and responsible GenAI adoption.

G. Towards Sustainable and Responsible GenAI Integration

Overall, the findings suggest that the integration of GenAI in programming education should move beyond a purely performance-driven approach toward a more responsibility-oriented framework. While GenAI demonstrates clear benefits in enhancing learning outcomes, its long-term value depends on how responsibly it is adopted and managed.

Sustainable integration requires a careful balance between technological capability, pedagogical alignment, ethical governance, and institutional readiness. These dimensions must work together to ensure that GenAI supports meaningful learning while preserving academic integrity, critical thinking, and educational quality.

VI. Conclusion and Future Work

A. Conclusion

Generative Artificial Intelligence (GenAI) has emerged as a significant development in higher education, with particular relevance for programming and computer science education. Its growing use in educational settings has introduced new opportunities for code generation, real-time feedback, conceptual explanation, and adaptive learning support. At the same time, its integration also raises important pedagogical, ethical, and institutional challenges that are not yet fully understood.

This systematic literature review synthesised evidence from 60 empirical studies published between 2022 and 2025, providing a structured overview of GenAI use in higher education, with particular emphasis on programming and computer science education. In relation to RQ1, the findings show that GenAI can positively influence learning outcomes by improving code correctness, debugging efficiency, conceptual understanding, and student engagement. These results suggest that GenAI functions not only as a productivity tool but also as a learning support system that can enhance the educational process.

Addressing RQ2, the review reveals considerable variation in GenAI adoption across countries, institutions, course contexts, and access conditions. These differences are shaped by disparities in infrastructure, policy frameworks, institutional readiness, and resource availability, highlighting the importance of context in shaping effective adoption. In response to RQ3, the findings indicate that both students and lecturers generally perceive GenAI positively. However, continued use is influenced by more than usefulness alone. Factors such as satisfaction, trust, motivation, and ethical alignment play an important role in shaping sustained engagement. Regarding RQ4, the review identifies major challenges related to academic integrity, over-reliance, reliability, and ethical misuse, underscoring the need for careful pedagogical design, clear institutional policies, and responsible governance.

Taken together, these findings highlight the need for a more integrated, theory-informed, and pedagogically grounded approach to GenAI adoption. Such an approach should align technological capability with educational goals, ethical responsibility, and institutional readiness in order to support more sustainable and meaningful use of GenAI in higher education, particularly in programming and computer science education.

B. Future Work

Based on the identified gaps and limitations in the literature, this review proposes several directions for future research, practice, and policy development:

1) Longitudinal and experimental evaluation: Future research should move beyond short-term performance measures to examine the long-term impact of GenAI on learning, cognitive development, problem-solving ability, and computational thinking. Such studies are necessary to determine whether the observed benefits of GenAI translate into sustained educational outcomes over time.

2) Expansion into advanced programming and broader disciplinary contexts: Further studies are needed to explore the role of GenAI in more advanced areas of programming and computer science education, including algorithms, software engineering, and complex system design. At the same time, broader higher education contexts should also be examined to understand how GenAI functions across disciplines with different learning demands.

3) Addressing global and institutional disparities: More research is needed to investigate how differences in infrastructure, policy frameworks, access conditions, and institutional readiness shape equitable adoption across regions, particularly in under-resourced and developing contexts.

4) Pedagogical integration and instructor role: Future work should examine how GenAI can be effectively embedded within teaching practice, including curriculum redesign,

instructor training, and the development of AI-supported strategies that promote responsible and meaningful learning.

5) Evaluation of governance and academic integrity frameworks: There is a need for empirical research that evaluates the effectiveness of institutional policies, ethical guidelines, and assessment redesign strategies in addressing academic integrity concerns and reducing misuse.

6) Theory-driven and interdisciplinary approaches: Future research should adopt stronger theoretical foundations by extending models such as ECM, TAM, and UTAUT, while integrating perspectives from educational theory, human–computer interaction, and ethics to develop more comprehensive and context-sensitive frameworks for GenAI in higher education.

7) Practical implications for educators and policymakers: For educators, the findings suggest the need to integrate GenAI through guided instructional design, AI literacy activities, assessment redesign, and critical evaluation of AI-generated outputs. For policymakers and institutional leaders, the results highlight the importance of clear academic integrity policies, equitable access to GenAI tools, staff training, ethical governance, and continuous monitoring of GenAI use in teaching and learning. These measures are essential to ensure that GenAI supports meaningful, inclusive, and responsible education.

The rapid evolution of Generative Artificial Intelligence (GenAI) is reshaping teaching and learning in higher education, with particularly important implications for programming and computer science education. Ensuring its effective and responsible integration requires more than technological advancement; it demands a coordinated effort to align pedagogy, ethical principles, and institutional governance. As this review has shown, the foundation for GenAI-supported learning has been established, but it remains incomplete and still evolving. The next phase of research and practice should therefore focus on developing sustainable, inclusive, and theory-informed frameworks that position GenAI not merely as a tool for efficiency, but as a resource for meaningful, equitable, and responsible education.

REFERENCES

[1] D. Lee et al., "The impact of generative AI on higher education learning and teaching: A study of educators' perspectives," Comput. Educ. Artif. Intell., vol. 6, p. 100221, Jun. 2024, doi: 10.1016/j.caeai.2024.100221.

[2] D. Kohen-Vacs, M. Usher, and M. Jansen, "Integrating Generative AI into Programming Education: Student Perceptions and the Challenge of Correcting AI Errors," Int. J. Artif. Intell. Educ., Jul. 2025, doi: 10.1007/s40593-025-00496-4.

[3] J. Llerena-Izquierdo, J. Mendez-Reyes, R. Ayala-Carabajo, and C. Andrade-Martinez, "Innovations in Introductory Programming Education: The Role of AI with Google Colab and Gemini," Educ. Sci., vol. 14, no. 12, p. 1330, Dec. 2024, doi: 10.3390/educsci14121330.

[4] R. Mellado, C. Cubillos, and G. Ahumada, "Effectiveness of Generative Artificial Intelligence in learning programming to higher education students," in 2024 IEEE International Conference on Automation/XXVI Congress of the Chilean Association of Automatic Control (ICA-ACCA), Santiago, Chile: IEEE, Oct. 2024, pp. 1–7. doi: 10.1109/ICA-ACCA62622.2024.10766746.

[5] R. Yilmaz and F. G. Karaoglan Yilmaz, "The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation," Comput. Educ. Artif. Intell., vol. 4, p. 100147, 2023, doi: 10.1016/j.caeai.2023.100147.

[6] A. A. Hernandez, E. M. Albina, and A. R. Caballero, "Generative Artificial Intelligence for Programming Education: Enhancing Input to Outcome-Based Teaching and Learning Approaches," in 2025 Seventh International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan: IEEE, Jun. 2025, pp. 1–5. doi: 10.1109/IS3C65361.2025.11131023.

[7] G. Akçapınar and E. Sidan, "AI chatbots in programming education: guiding success or encouraging plagiarism," Discov. Artif. Intell., vol. 4, no. 1, p. 87, Nov. 2024, doi: 10.1007/s44163-024-00203-7.

[8] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: perceptions, benefits, and challenges in higher education," Int. J. Educ. Technol. High. Educ., vol. 20, no. 1, p. 43, Jul. 2023, doi: 10.1186/s41239-023-00411-8.

[9] H. Wang, A. Dang, Z. Wu, and S. Mac, "Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines," Comput. Educ. Artif. Intell., vol. 7, p. 100326, Dec. 2024, doi: 10.1016/j.caeai.2024.100326.

[10] K. Aruleba, I. T. Sanusi, G. Obaido, B. Ogbuokiri, and I. D. Mienye, "Beyond the Prompt: Student Strategies, Ethical Reflections, and Learning with ChatGPT in Computer Science," Technol. Knowl. Learn., Aug. 2025, doi: 10.1007/s10758-025-09899-7.

[11] M. I. Baig and E. Yadegaridehkordi, "Factors influencing academic staff satisfaction and continuous usage of generative artificial intelligence (GenAI) in higher education," Int. J. Educ. Technol. High. Educ., vol. 22, no. 1, Feb. 2025, doi: 10.1186/s41239-025-00506-4.

[12] M. Bikanga Ada, "It Helps with Crap Lecturers and Their Low Effort: Investigating Computer Science Students' Perceptions of Using ChatGPT for Learning," Educ. Sci., vol. 14, no. 10, p. 1106, Oct. 2024, doi: 10.3390/educsci14101106.

[13] S. Bjelobaba, L. Waddington, M. Perkins, T. Foltýnek, S. Bhattacharyya, and D. Weber-Wulff, "Maintaining research integrity in the age of GenAI: an analysis of ethical challenges and recommendations to researchers," Int. J. Educ. Integr., vol. 21, no. 1, p. 18, Jul. 2025, doi: 10.1007/s40979-025-00191-w.

[14] V. Venkatesh, J. Y. Thong, and X. Xu, "Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead," J. Assoc. Inf. Syst., vol. 17, no. 5, pp. 328–376, 2016, doi: 10.17705/1jais.00428.

[15] M.-T. Huynh and T. Aichner, "In generative artificial intelligence we trust: unpacking determinants and outcomes for cognitive trust," AI Soc., May 2025, doi: 10.1007/s00146-025-02378-8.

[16] R. S. Pasupuleti and D. Thiyyagura, "An empirical evidence on the continuance and recommendation intention of ChatGPT among higher education students in India: An extended technology continuance theory," Educ. Inf. Technol., Mar. 2024, doi: 10.1007/s10639-024-12573-7.

[17] Y. Jin, L. Yan, V. Echeverria, D. Gašević, and R. Martinez-Maldonado, "Generative AI in higher education: A global perspective of institutional adoption policies and guidelines," Comput. Educ. Artif. Intell., vol. 8, p. 100348, Jun. 2025, doi: 10.1016/j.caeai.2024.100348.

[18] M.-L. Chiu, "Exploring user awareness and perceived usefulness of generative AI in higher education: The moderating role of trust," Educ. Inf. Technol., May 2025, doi: 10.1007/s10639-025-13612-7.

[19] C. Cho and D. Ofosu-Anim, "Exploring international students' perceptions of adopting generative artificial intelligence (GenAI) technologies in learning," Soc. Sci. Humanit. Open, vol. 11, p. 101418, 2025, doi: 10.1016/j.ssaho.2025.101418.

[20] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[21] G. W. Cheung, H. D. Cooper-Thomas, R. S. Lau, and L. C. Wang, "Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations," Asia Pac. J. Manag., vol. 41, no. 2, pp. 745–783, Jun. 2024, doi: 10.1007/s10490-023-09871-y.

[22] Y. Alagöz Hamzaj, "Generative AI acceptance among future educators: personality and behavioral insights," Educ. Inf. Technol., Jul. 2025, doi: 10.1007/s10639-025-13678-3.

[23] I. T. Awidi, "Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool," Comput. Educ. Artif. Intell., vol. 6, p. 100226, Jun. 2024, doi: 10.1016/j.caeai.2024.100226.

[24] A. L. Irish, M. W. Gazica, and V. Becerra, "A qualitative descriptive analysis on generative artificial intelligence: bridging the gap in pedagogy to prepare students for the workplace," Discov. Educ., vol. 4, no. 1, p. 48, Mar. 2025, doi: 10.1007/s44217-025-00435-4.

[25] S. Otto, S. Ejsing-Duun, and E. Lindsay, "Disruptive tensions and emerging practices: an exploratory inquiry into student perspectives on generative Artificial Intelligence in a problem-based learning environment," Educ. Inf. Technol., vol. 30, no. 13, pp. 19111–19140, Aug. 2025, doi: 10.1007/s10639-025-13533-5.

[26] M. N. Yakubu, N. David, and N. H. Abubakar, "Students' behavioural intention to use content generative AI for learning and research: A UTAUT theoretical perspective," Educ. Inf. Technol., vol. 30, no. 13, pp. 17969–17994, Aug. 2025, doi: 10.1007/s10639-025-13441-8.

[27] A. Yusuf, N. Pervin, and M. Román-González, "Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives," Int. J. Educ. Technol. High. Educ., vol. 21, no. 1, p. 21, Mar. 2024, doi: 10.1186/s41239-024-00453-6.

[28] C. K. Y. Chan and L. H. Y. Tsi, "Will generative AI replace teachers in higher education? A study of teacher and student perceptions," Stud. Educ. Eval., vol. 83, p. 101395, Dec. 2024, doi: 10.1016/j.stueduc.2024.101395.

[29] S. Nikolic et al., "ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering," Australas. J. Eng. Educ., vol. 29, no. 2, pp. 126–153, Jul. 2024, doi: 10.1080/22054952.2024.2372154.

[30] F. Deriba, I. T. Sanusi, O. O.Campbell, and S. S. Oyelere, "Computer Programming Education in the Age of Generative AI: Insights from Empirical Research," SSRN Electron. J., 2024, doi: 10.2139/ssrn.4891302.

[31] P. Suchanek and M. Kralova, "Generative artificial intelligence expectations and experiences in management education: ChatGPT use and student satisfaction," J. Innov. Knowl., vol. 10, no. 5, p. 100781, Sep. 2025, doi: 10.1016/j.jik.2025.100781.

[32] A. E. Sousa and P. Cardoso, "Use of Generative AI by Higher Education Students," Electronics, vol. 14, no. 7, p. 1258, Mar. 2025, doi: 10.3390/electronics14071258.

[33] X. Ren and M. L. Wu, "Examining Teaching Competencies and Challenges While Integrating Artificial Intelligence in Higher Education," TechTrends, vol. 69, no. 3, pp. 519–538, May 2025, doi: 10.1007/s11528-025-01055-3.

[34] S. C. Kong, Y. Yang, and C. Hou, "Examining teachers' behavioural intention of using generative artificial intelligence tools for teaching and learning based on the extended technology acceptance model," Comput. Educ. Artif. Intell., vol. 7, p. 100328, Dec. 2024, doi: 10.1016/j.caeai.2024.100328.

[35] S. Ivanov, M. Soliman, A. Tuomi, N. A. Alkathiri, and A. N. Al-Alawi, "Drivers of generative AI adoption in higher education through the lens of the Theory of Planned Behaviour," Technol. Soc., vol. 77, Jun. 2024, doi: 10.1016/j.techsoc.2024.102521.

[36] Y. Yang, Q. Xia, C. Liu, and T. K. F. Chiu, "The impact of TPACK on teachers' willingness to integrate generative artificial intelligence (GenAI): The moderating role of negative emotions and the buffering effects of need satisfaction," Teach. Teach. Educ., vol. 154, p. 104877, Feb. 2025, doi: 10.1016/j.tate.2024.104877.

[37] D. Tbaishat, G. Amoudi, and M. Elfadel, "Adapting teaching and learning with existing generative AI by higher education Students: Comparative study of Zayed University and King Abdulaziz University," Comput. Educ. Artif. Intell., vol. 8, p. 100421, Jun. 2025, doi: 10.1016/j.caeai.2025.100421.

[38] L. Kohnke, B. L. Moorhouse, and D. Zou, "Exploring generative artificial intelligence preparedness among university language instructors: A case study," Comput. Educ. Artif. Intell., vol. 5, p. 100156, 2023, doi: 10.1016/j.caeai.2023.100156.

[39] T. K. F. Chiu, "Future research recommendations for transforming higher education with generative AI," Computers and Education: Artificial Intelligence, vol. 6. Elsevier B.V., 2024. doi: 10.1016/j.caeai.2023.100197.

[40] L. Kohnke, "Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning," Comput. Educ. Artif. Intell., vol. 7, p. 100279, Dec. 2024, doi: 10.1016/j.caeai.2024.100279.

[41] M. B. Ulla, M. J. C. Advincula, C. D. S. Mombay, H. M. A. Mercullo, J. P. Nacionales, and A. D. Entino-Señorita, "How can GenAI foster an inclusive language classroom? A critical language pedagogy perspective from Philippine university teachers," Comput. Educ. Artif. Intell., vol. 7, p. 100314, Dec. 2024, doi: 10.1016/j.caeai.2024.100314.

[42] D. Huang, N. Hash, J. J. Cummings, and K. Prena, "Academic cheating with generative AI: Exploring a moral extension of the theory of planned behavior," Comput. Educ. Artif. Intell., vol. 8, p. 100424, Jun. 2025, doi: 10.1016/j.caeai.2025.100424.

[43] N. Hashmi and A. S. Bal, "Generative AI in higher education and beyond," Bus. Horiz., vol. 67, no. 5, pp. 607–614, Sep. 2024, doi: 10.1016/j.bushor.2024.05.005.

[44] K. A. Charles, A. Yousuf, H. C. Chua, S. Matthews, J. Harnett, and T. Hinton, "AI in action: Changes to student perceptions when using generative artificial intelligence for the creation of a multimedia project-based assessment," Eur. J. Pharmacol., vol. 998, p. 177508, Jul. 2025, doi: 10.1016/j.ejphar.2025.177508.

[45] E. Ukwandu, O. Omisade, K. Jones, S. Thorne, and M. Castle, "The Future of Teaching and Learning In The Context Of Emerging Artificial Intelligence Technologies," 2024, SSRN. doi: 10.2139/ssrn.4852030.

[46] B. Zou, C. Wang, H. He, C. Li, E. Purwanto, and P. Wang, "Enhancing EFL writing with visualised GenAI feedback: A cognitive affective theory of learning perspective on revision quality, emotional response, and human-computer interaction," Learn. Motiv., vol. 91, p. 102158, Aug. 2025, doi: 10.1016/j.lmot.2025.102158.

[47] N. N. Chan et al., "Generative artificial intelligence in a VUCA world: the 'Lived Experiences' of Southeast Asian teachers' use of AI in higher education," Int. J. Educ. Res., vol. 133, p. 102733, 2025, doi: 10.1016/j.ijer.2025.102733.

[48] G. Gmiterek and S. D. Kotuła, "Generative artificial intelligence in the activities of academic libraries of public universities in Poland," J. Acad. Librariansh., vol. 51, no. 3, p. 103043, May 2025, doi: 10.1016/j.acalib.2025.103043.

[49] K. Beckman, T. Apps, S. K. Howard, C. Rogerson, A. Rogerson, and J. Tondeur, "The GenAI divide among university students: A call for action," Internet High. Educ., vol. 67, p. 101036, Oct. 2025, doi: 10.1016/j.iheduc.2025.101036.

[50] H. Zhao, B. Yuan, and Y. Song, "Employees' perception of generative artificial intelligence and the dark side of work outcomes," J. Hosp. Tour. Manag., vol. 61, pp. 191–199, Dec. 2024, doi: 10.1016/j.jhtm.2024.10.007.

[51] V. Ratten and P. Jones, "Generative artificial intelligence (ChatGPT): Implications for management educators," Int. J. Manag. Educ., vol. 21, no. 3, p. 100857, Nov. 2023, doi: 10.1016/j.ijme.2023.100857.

[52] M. Ostick, B. Mariani, and C. Lovecchio, "The use of generative artificial intelligence (AI) in nursing education," Teach. Learn. Nurs., p. S1557308725001799, Jun. 2025, doi: 10.1016/j.teln.2025.05.025.

[53] Y. Dai, A. Liu, and C. P. Lim, "Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education," Procedia CIRP, vol. 119, pp. 84–90, 2023, doi: 10.1016/j.procir.2023.05.002.

[54] Y. Cha, Y. Dai, Z. Lin, A. Liu, and C. P. Lim, "Empowering University Educators to Support Generative AI-enabled Learning: Proposing a Competency Framework," Procedia CIRP, vol. 128, pp. 256–261, 2024, doi: 10.1016/j.procir.2024.06.021.

[55] L. G. A. De Putter-Smits, C. F. J. Pols, P. J. J. M. Dekkers, P. R. Runhaar, M. Timmer, and J. T. Van Der Veen, "Exploring the role of generative AI in science teacher education programs: a qualitative study," Int. J. Educ.

Res. Open, vol. 9, p. 100492, Dec. 2025, doi: 10.1016/j.ijedro.2025.100492.

[56] N. McDonald, A. Johri, A. Ali, and A. H. Collier, "Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines," Comput. Hum. Behav. Artif. Hum., vol. 3, p. 100121, Mar. 2025, doi: 10.1016/j.chbah.2025.100121.

[57] V. C. Chan, "Integrating generative artificial intelligence in a writing intensive course for undergraduate nursing students," J. Prof. Nurs., vol. 57, pp. 85–91, Mar. 2025, doi: 10.1016/j.profnurs.2025.01.003.

[58] S. M. Sarsam, A. A. Alias, C. S. Mon, H. Al-Samarraie, and A. I. Al-Hatem, "Exploring Public Opinions Toward the Use of Generative Artificial Intelligence Chatbot in Higher Education: An Insight from Topic Modelling and Sentiment Analysis," in 2023 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria: IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/BdKCSE59280.2023.10339760.

[59] S. Otto, R. Lavi, and L. Brogaard Bertel, "Human-GenAI interaction for active learning in STEM education: State-of-the-art and future directions," Comput. Educ., vol. 239, p. 105444, Dec. 2025, doi: 10.1016/j.compedu.2025.105444.

[60] S. M. Nicolajsen, S. Nielsen, L. M. Carlsen, and C. Brabrand, "Programming education across disciplines: a nationwide study of Danish higher education," High. Educ., Dec. 2024, doi: 10.1007/s10734-024-01345-4.

[61] S. Cengiz and A. Peker, "Generative artificial intelligence acceptance and artificial intelligence anxiety among university students: the sequential mediating role of attitudes toward artificial intelligence and literacy," Curr. Psychol., vol. 44, no. 9, pp. 7991–8000, May 2025, doi: 10.1007/s12144-025-07433-7.

[62] Y. Qu, M. X. Y. Tan, and J. Wang, "Disciplinary differences in undergraduate students' engagement with generative artificial intelligence," Smart Learn. Environ., vol. 11, no. 1, p. 51, Nov. 2024, doi: 10.1186/s40561-024-00341-6.

[63] J. Heil, D. Ifenthaler, M. Cooper, M. L. Mascia, R. Conti, and M. P. Penna, "Students' perceived impact of GenAI tools on learning and assessment in higher education: the role of individual AI competence," Smart Learn. Environ., vol. 12, no. 1, p. 37, May 2025, doi: 10.1186/s40561-025-00395-0.

[64] Y. An, J. H. Yu, and S. James, "Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration," Int. J. Educ. Technol. High. Educ., vol. 22, no. 1, p. 10, Feb. 2025, doi: 10.1186/s41239-025-00507-3.

[65] K. Zdravkova and B. Ilijoski, "The impact of large language models on computer science student writing," Int. J. Educ. Technol. High. Educ., vol. 22, no. 1, p. 32, May 2025, doi: 10.1186/s41239-025-00525-1.

[66] Ö. F. Ursavaş, Y. Yalçın, H. İslamoğlu, E. Bakır-Yalçın, and M. Cukurova, "Rethinking the importance of social norms in generative AI adoption: investigating the acceptance and use of generative AI among higher education students," Int. J. Educ. Technol. High. Educ., vol. 22, no. 1, p. 38, Jun. 2025, doi: 10.1186/s41239-025-00535-z.

[67] L. Fan, K. Deng, and F. Liu, "Educational impacts of generative artificial intelligence on learning and performance of engineering students in China," Sci. Rep., vol. 15, no. 1, p. 26521, Jul. 2025, doi: 10.1038/s41598-025-06930-w.

[68] A. Dabis and C. Csáki, "AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI," Humanit. Soc. Sci. Commun., vol. 11, no. 1, p. 1006, Aug. 2024, doi: 10.1057/s41599-024-03526-z.

Translating Job Advertisements into Competency Taxonomy: An Interpretable Approach for Robotics Recruitment Analysis

Zhiyan Xue[1], Yang Zhou[2]*
School of Electronic Information Engineering, Suzhou Polytechnic University, Suzhou, China [1]
School of Management, Suzhou Polytechnic University, Suzhou, China [2]

***Abstract*—To translate robotics recruitment texts into an interpretable competency taxonomy, this study explored how latent topics can be induced from large-scale job advertisements and used to construct a structured competency taxonomy. A domain-specific preprocessing pipeline was applied to a corpus of robotics job advertisements, and latent topics were subsequently induced using Latent Dirichlet Allocation (LDA). Building on the extracted topic evidence, a procedure was applied to organize topic summaries into task domains and competencies grounded in topic keywords. Evaluation was conducted using a set of quantitative measures to assess semantic consistency and structural quality. The results revealed recurring competency patterns encompassing on-site operation and service, engineering design and integration, software and algorithm development, and system verification and reliability assurance. The resulting competency taxonomy captures the underlying structure of employer demand and provides an interpretable basis for robotics skill demand analysis, supporting role profiling and serving as an empirical reference for Vocational Education and Training (VET).**

Keywords*—Competency taxonomy; robotics industry; Latent Dirichlet Allocation (LDA); Vocational Education and Training (VET)

I. Introduction

A. Research Background

Robotics has expanded from laboratory research into broad industrial and service deployments, including manufacturing, logistics, healthcare, and consumer-facing applications. This expansion has diversified robotics job roles and increased demand for talent with cross-disciplinary skill profiles. At the same time, rapid iteration of toolchains and deployment practices shortens the validity period of static skill descriptions. These trends create a practical need for methods that can characterize robotics competency requirements in a scalable and updatable manner [1][2].

Online job advertisements provide a timely and large-scale source of evidence for this purpose. Recruitment texts serve as explicit, employer-authored descriptions of hiring requirements and are typically high in volume, frequently updated, and broad in coverage. As a result, job advertisements can be treated as a near real-time signal of skill demand, enabling analysis of the structure of required competencies and their evolution across job families and time windows [3][4]. Compared with interviews or small-sample approaches, recruitment texts support repeatable analysis at scale, which is particularly valuable in fast-changing technical domains [5][6]. Robotics recruitment introduces additional complexity that makes competency modeling nontrivial. Robotics roles commonly combine capabilities spanning electromechanical systems, software control, as well as operation and maintenance. Skill evidence is distributed across a chain of job families that involve mechanical design, electrical engineering, embedded development, systems integration, and on-site support. Consequently, skill expressions in job advertisements are fragmented, terminology is not standardized across enterprises, and the granularity of requirements varies from broad capability statements to tool-specific and task-level descriptions [7][8]. Traditional interview-based or small-sample approaches often fail to yield a competency structure that generalizes across diverse job roles [9][10][11].

Against this background, two research questions guide the present work. What latent competency domains and skill patterns are reflected in robotics recruitment texts? How can unsupervised topic evidence from recruitment texts be translated into an interpretable and well-structured competency taxonomy?

To address these two questions, an approach is proposed that translates job advertisements into a structured competency taxonomy. The approach proceeds from recruitment text preprocessing to LDA-based topic induction, followed by translation from topic summaries to task domains and competencies grounded in topic keywords. Evaluation is designed for unlabeled settings and emphasizes interpretability and structural validity. Topic coherence is used to assess the semantic consistency of topic summaries, and a set of quantitative measures is used to quantify representativeness, noise, and redundancy of the induced taxonomy. Together, these components provide an interpretable basis for robotics skill demand mining from job advertisements, while supporting comparison against standard baselines in applied text mining.

B. Related Work

With the rapid development of the robotics industry, Vocational Education and Training (VET) systems have been under sustained pressure to update talent cultivation promptly. Existing studies have shown that educational and training responses often lag behind technological progress and shifts in market demand. Accordingly, scholars have proposed a range of strategies at different levels to improve talent development

*Corresponding author.

in robotics-related fields [12][13][14][15]. Existing studies have mainly approached this issue from two closely related directions. The first strand focuses on curriculum systems and talent cultivation models. Studies in this line of research primarily examine curriculum design, training mechanisms, and pathways for industry-education integration in robotics-related programs to enhance the alignment between educational provision and industrial demand. For example, Wang et al. pointed out that school-enterprise cooperation is an effective approach to improving teaching quality, thereby highlighting industrial demand as an important reference for curriculum design and educational improvement [12].

The second strand concentrates on specific skill demands and occupational competency requirements. This line of research mainly analyzes particular technical domains or concrete occupational contexts. Specifically, Yu-Shen et al. [13] focused on specific skill demands in the robotics field; Olukanni et al. [14] further situated the analysis in concrete work settings, such as human-robot collaboration, and Do et al. [15] through an Importance-Performance Analysis (IPA), suggested that university curricula should place greater emphasis on courses such as Artificial Intelligence, Machine Vision, and Robot Structure Design. These studies have helped identify key competency dimensions required in robotics-related occupations and have provided relatively direct practical implications for the optimization of course content.

Overall, existing research on robotics talent cultivation and competency development has largely proceeded from the supply side, with emphasis on educational planning and curriculum reform. While such studies are valuable for defining training objectives and competency expectations, they provide only limited insight into how firms specify skill requirements in actual recruitment and how these requirements evolve with industrial change. In parallel, studies addressing specific robotics skills or competency dimensions often focus on a single technical direction or a limited set of occupational contexts. Sample scopes are therefore typically constrained, and the resulting outputs frequently take the form of partial skill lists tied to local tasks or particular positions. Such a pattern makes it difficult to cover the major upstream and downstream positions across the robotics industrial chain.

Beyond the substantive findings of robotics-specific studies, a related body of research has approached the issue from a methodological perspective by examining how occupational skills can be identified, constructed, and organized. From this perspective, existing studies can be broadly grouped into three methodological approaches. The first perspective relies on expert-based methods, such as Delphi studies, expert interviews, and literature synthesis [15][16]. A notable strength of this line of work is that the resulting frameworks are usually grounded in expert understanding of occupational practice and therefore tend to remain highly interpretable and closely aligned with domain knowledge [17]. However, such methods are often costly to update and dependent on expert judgment. Consequently, they are well suited to normative framework construction, but less suited to tracking the changing skill requirements of large numbers of job advertisements. The second perspective uses recruitment texts as a demand evidence source for skill analysis. Online job advertisements can be treated as employer demand statements that contain explicit signals about skills, work activities, and job requirements [18][19]. Some studies have used rule extraction and statistical representation tools to identify terms from job advertisements [20]. Findings from this stream of research indicate that recruitment texts can support large-scale labor market and skill analysis and can reveal meaningful patterns in employer demand. However, the resulting outputs often remain fragmented because they are typically expressed as keyword lists. The third perspective extends recruitment text analysis toward the discovery of latent structure. Topic modeling, including LDA and other related approaches, has been used to identify recurring themes [21]. In recruitment analytics, these methods have been applied to domains such as big data analytics and software engineering to derive topic-based representations of knowledge domains, skill combinations [22][23][24]. Compared with direct term extraction, topic modeling provides a more effective basis for organizing dispersed skill expressions into coherent thematic structures. Nevertheless, a topic provides a meaningful set of salient terms, but it does not explain how those terms should be translated into hierarchical competency domains [25]. In other words, existing recruitment text studies have shown that skill signals can be extracted and that latent themes can be discovered, yet they have not fully resolved how topic-level evidence should be converted into a competency structure.

Building on the foregoing methodological discussion, a further challenge concerns how the outcomes of skill and competency induction should be evaluated. This issue has received relatively limited attention, largely because competency extraction from recruitment text is often conducted in the absence of standardized labels or benchmark datasets [26]. Soare [27] and McClelland [28] relied on expert-curated competency frameworks and skill taxonomies as references for workforce planning and training design, which provided guidance on how competency structures could be organized but did not yield direct evaluation criteria. Hoyle et al. [29] adopted topic coherence as an intrinsic indicator of semantic consistency for topic word sets, which offered a practical way to judge whether a topic summary formed a coherent group of terms. Kim et al. [30] further suggested that representativeness and distinctiveness were essential for reusable and computationally meaningful competency structures, which implied that coherence alone was insufficient because it did not assess document-level representativeness or cross-topic redundancy. This line of work motivated evaluation protocols that complement coherence with structure-oriented indicators, yet a unified evaluation approach for determining whether coherent topic summaries translate into verifiable and non-redundant competency units remains absent.

The above literature highlighted two gaps that motivated this study. First, in the robotics domain, existing studies on talent development and skill requirements have largely proceeded from the supply side and have not yet established a demand-side competency structure grounded in large-scale recruitment evidence across the major positions of the robotics industry. Second, although existing recruitment text studies have identified skill signals and uncovered latent topics, these outputs have not formed a competency taxonomy. More

importantly, approaches for translating topic evidence into an interpretable competency structure have remained underdeveloped. This study addressed these gaps by introducing a translation approach and evaluation metrics to enable interpretable competency induction in robotics recruitment analysis.

The structure of this study is organized as follows. Section II describes the method, including data collection, preprocessing, topic discovery using LDA, and translation from topic summaries to a competency taxonomy. Section III introduces the experimental setup, baseline, and evaluation metrics. Section IV reports results, including both quantitative findings and qualitative case analyses. Section V concludes with limitations and future directions.

II. Methodology

This study proposed a structured pipeline that transforms unstructured job advertisements into an interpretable competency taxonomy. The process consists of five core stages: 1) Data Collection; 2) Data Preparation; 3) Topic Discovery; 4) Topic-to-Competency Translation; 5) Quantitative Evaluation. This pipeline takes raw job recruitment texts as input and produces a competency taxonomy as output. Fig. 1 illustrates the overall framework.

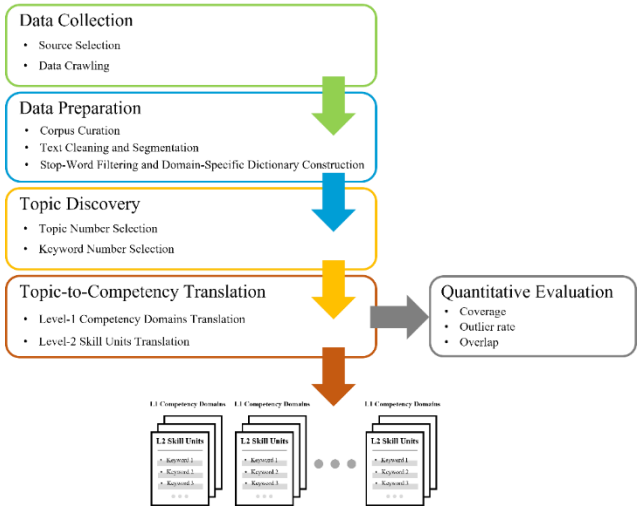


Fig. 1. The proposed study framework.

A. Data Collection

1) Source selection: Robotics job advertisements are distributed across multiple online platforms and often exhibit inconsistent formatting and varying levels of detail. To balance market representativeness with textual suitability for systematic analysis, the study selects Zhaopin (www.zhaopin.com) as the data source for three considerations. First, Zhaopin is one of the leading online recruitment platforms in China, with broad geographic coverage and substantial posting volume. Its market presence facilitates the collection of a diverse sample of robotics-related positions and helps reduce potential selection bias. Second, job advertisements on Zhaopin tend to follow a relatively standardized structure and are subject to platform-level quality control, which makes them well suited for natural language processing (NLP). In particular, the platform encourages employers to provide detailed job description texts that explicitly separate "position responsibilities" and "job requirements", offering a high-quality corpus for extracting skill-related signals. Third, Zhaopin data have been used in prior peer-reviewed studies to analyze labor market demand patterns, providing an external validity signal for leveraging its job description texts in skill demand mining [31][32]. To reduce temporal drift and ensure that the induced topics reflect a consistent market snapshot, the study restricts data collection to a fixed sampling window spanning December 2025 to January 2026.

2) Data collection: Selenium was adopted as the primary tool for crawling and parsing job advertisement pages in this phase. It is a widely used automation framework that programmatically controls real web browsers, enabling reliable interaction with web pages [33][34]. To target robotics-related vacancies while maintaining regional diversity, the study queried the platform using the keywords "Robotics" and "Robot", which aim to capture openings across the robotics value chain, including R&D, system integration, and maintenance. Search locations were restricted to six major regions in China: Beijing, Jiangsu, Shanghai, Zhejiang, Guangdong, and Shandong Province. This location set was chosen because these regions collectively represent a large share of the national robotics industrial chain and host a high concentration of industrial robot manufacturing enterprises. For each location, the crawler iterated over search result pages, collected posting URLs, and parsed the corresponding detail pages to extract the job title, full job description text, and other data such as company name and salary. The initial crawl yielded 6,802 advertisements, reflecting the current labor demand for robotics-related positions in the aforementioned cities.

B. Data Preparation

1) Corpus curation: At this stage, the study curated a corpus of valid robotics-related positions to reduce noise that might interfere with downstream topic discovery. The advertisements were curated through three steps: removing invalid or incomplete records, excluding irrelevant advertisements such as finance jobs, administrative jobs, and promotional or advertisement entries, and eliminating exact duplicate advertisements. After cleaning and deduplication, the resulting corpus contains 3,290 job descriptions and serves as the input to subsequent preprocessing.

2) Text cleaning and segmentation: Recruitment texts often contain punctuation noise, irregular symbols, and platform-specific artifacts that are not informative for skill demand mining. A cleaning procedure is performed before segmentation. Specifically, Python regular expressions are applied via re.sub() to remove punctuation marks and abnormal characters and to normalize whitespace. Additionally, unlike English processing pipelines, Chinese

text does not require uppercase or lowercase conversion, lemmatization, or stemming; therefore, emphasis is placed on accurate word segmentation and term normalization as the primary preprocessing steps.

The corpus is then segmented using Jieba, a widely used Chinese word segmentation toolkit that supports dictionary-based tokenization through prefix-dictionary matching and statistical disambiguation. Jieba is selected for its robustness on large-scale web text, its ability to incorporate a lexicon to preserve robotics technical terms and vendor names, and its lightweight, reproducible integration in standard Python environments, which facilitates transparent replication of the preprocessing pipeline [35].

3) Stop-word filtering and domain-specific dictionary construction: Recruitment corpora contain a substantial amount of platform-specific and HR-related expressions. To mitigate this effect, stop-word filtering is performed using an expanded stop-word set based on the widely used Harbin Institute of Technology stop-word dictionary (hit_stopwords), which is commonly adopted in Chinese NLP studies due to its broad coverage of high-frequency non-content terms [36]. Building on this foundation, the stop-word set is further extended with recruitment-related terms that are prevalent in job advertisements but provide limited value for skill demand analysis. Typical examples include template headers and HR-related expressions such as “compensation and benefits”, “job responsibilities”, “job requirements/qualifications”, and “bonus points/plus”. Removing such terms reduces the risk that topic discovery is dominated by generic recruitment language rather than domain skills.

In addition to stop-word filtering, a domain-specific user dictionary is constructed to preserve robotics-related terms that might otherwise be incorrectly split or discarded. The dictionary covers two major categories. Robotics brands and company names are included, such as Siemens, KUKA, ABB, and FANUC. Technical terms are added to capture core skills, including key algorithms and platforms such as SLAM, ROS, OpenCV, and multimodal learning, as well as industrial automation and engineering terms such as CAD, AGV, Robot Studio, and robotic arm.

By integrating this user dictionary into the segmentation process, domain terms are more consistently tokenized as single units, reducing vocabulary sparsity and improving the stability and interpretability of the downstream topics and competency taxonomy.

C. Topic Discovery

LDA was employed to uncover latent competency topics from the robotics recruitment texts. Its selection was primarily guided by two considerations. First, the purpose of the study was to explore latent competency structures embedded in recruitment texts rather than to assign documents to predefined categories. Under this objective, an unsupervised topic modeling method was methodologically appropriate [37]. Second, given the scale of the corpus and the practical constraints on computational resources, LDA offered advantages in terms of computational efficiency and implementation feasibility. For these reasons, LDA was adopted in the present study. LDA represents each document as a mixture of latent topics, where each topic is modeled as a probability distribution over the vocabulary, thereby facilitating the extraction of interpretable themes representing skill domains and recurring requirements. Given K topics, the model estimates a topic-word distribution $\emptyset_K$ for each topic t_k ($k = 1, \dots, K$) and a document-topic distribution θ_i for each document d_i ($i = 1, \dots, N$) . These distributions provide a ranked list of words that characterize each topic.

Model training is implemented using the Scikit-learn (sklearn) module. Scikit-learn is a widely used open-source machine learning library in Python that provides consistent APIs for data preprocessing, model training, and evaluation across a broad range of algorithms, including topic models. Its LDA implementation offers practical advantages for applied studies, such as integration with standard pipelines and built-in metrics.

1) Topic number selection: In practice, the topic number K is determined by jointly considering topic coherence and perplexity to balance interpretability and model fit. This criterion helps mitigate the risk of underfitting (overly coarse topics) and overfitting (overly fragmented topics), yielding topics that remain interpretable while retaining reasonable explanatory power on data.

2) Keyword number selection: For each topic t_k, the Top-M keywords with the highest probabilities are extracted as an interpretable topic summary, which serves as a compact and easily understandable representation of the topic and is also used as the basis for downstream topic-to-competency translation and structural evaluation.

M is set to 30 as a balanced choice that trades off information coverage against noise introduction. When M is too small, secondary but still important skill terms may be omitted, causing topics to be under-expressed and reducing representational adequacy for recruitment competencies. When M is too large, long-tail and weakly related words are more likely to enter the summary, which can increase overlap between topics and weaken discriminability between topics.

D. Topic-to-Competency Translation

Topic keywords provide concise summaries, but do not directly constitute a competency taxonomy. To bridge this gap, each discovered topic is translated into a two-level structure consisting of task domains and competencies. The translation aims to preserve the semantic content of each topic while producing a structured representation that remains interpretable, traceable, and suitable for subsequent evaluation.

The translation is conducted independently for each topic, based on its ranked Top-M keyword list. No terms outside the current topic are introduced during translation. This constraint preserves topic discriminability and avoids artificial gains in coverage caused by borrowing terms from other topics. In addition, the procedure prioritizes representativeness, meaning that the resulting competency structure is expected to retain the dominant semantic content of the topic rather than rely on a

small number of isolated keywords. Redundancy control is also enforced, such that different competencies under the same task domain capture distinct functional elements rather than repeated variants of the same meaning. The procedure consists of five steps.

Step 1: Assign the task domain. The ranked Top-M keyword list of each topic is first examined as a whole in order to identify its dominant capability orientation. The task domain is then assigned as a task domain-level descriptor that summarizes the central functional area implied by the topic keywords. This naming process follows established robotics and engineering terminology whenever possible, so that the resulting label remains professionally interpretable and consistent with domain usage. The task domain label is required to reflect the main semantic center of the topic rather than a peripheral or overly generic aspect of the keyword set.

Step 2: Partition keywords into candidate competency subsets. After the task domain is determined, the same Top-M keyword list is partitioned into several candidate competency subsets representing fine-grained functional elements. Each subset is intended to capture one operationally meaningful competency, such as design, programming, debugging, testing, deployment, or maintenance-related capability. The grouping process remains restricted to the current topic and organizes keywords according to semantic coherence and task domain relevance. Composite or overly broad groupings are avoided when the underlying competencies can be separated into more precise units.

Step 3: Refine competency boundaries. The candidate competency subsets are then refined by prioritizing higher-ranked keywords, as these terms carry the strongest evidence of the semantic core of the task domain. Particular attention is paid to preserving the dominant task domain-level signals, rather than allowing the translation to be driven by lower-ranked or weakly informative terms. During this step, the boundaries between competencies are adjusted so that each competency retains a clear functional focus within the task domain. When two competencies exhibit substantial semantic overlap, they are merged or revised. When a competency combines multiple distinct functional elements, it is further divided into more fine-grained competencies.

Step 4: Check evidence retention and reduce redundancy. The provisional competency-level structure is then evaluated against the full Top-M keyword list of the topic. This audit examines whether the main topic evidence is retained, whether important high-ranking keywords remain unassigned to the induced structure, and whether different competencies reuse the same semantic content to an excessive degree. If substantial unassigned keywords are identified, especially among the higher-ranked terms, the partition is revised to improve representativeness. If strong overlap is observed between competencies, the structure is adjusted to improve non-redundancy. This step ensures that the translation does not become a purely naming process detached from the original topic evidence.

Step 5: Finalize Competency labels. Once the keyword partition reaches an acceptable balance between representativeness and redundancy, each subset is formalized as a competency. The corresponding competency label is expressed as a concise operational capability statement grounded in the assigned keyword subset. The final output for each topic therefore consists of one task domain, several competencies, and an explicit mapping from each competency to its supporting topic-specific keywords. This mapping preserves traceability between the induced competency structure and the original topic summary, thereby enabling subsequent audit of evidence retention and structural consistency.

Through this procedure, topic summaries are transformed into a two-level competency taxonomy. The resulting structure does not aim to establish a complete ontology of robotics competencies. Instead, it organizes the underlying topic evidence into a two-level representation consisting of task domains and competencies under unlabeled conditions.

E. Quantitative Evaluation

Quantitative evaluation is conducted to assess the interpretability and structural validity of topic summaries in an unsupervised setting without standardized labels. The evaluation protocol is based on the Top-M keyword summary of each topic. Specifically, three metrics are considered: coverage, outlier rate, and overlap.

1) Coverage: Coverage evaluates whether the Top-M keyword summary is representative of the documents assigned to a topic. Let C_k denote the set of documents associated with topic k, and let T_k denote its Top-M keyword set. Let $W(d)$ be the set of tokens appearing in document d. Document-level coverage for topic k is defined as:

$$\mathrm{Cov}(k) = \frac{1}{|C_k|}\sum_{d\in C_k} I\left(W(d)\cap T_k \neq \emptyset\right) \quad (1)$$

where, $I(\cdot)$ is the indicator function. The corpus-level coverage is computed as a weighted average:

$$\mathrm{Cov} = \sum_{k=1}^{K}\frac{|C_k|}{N}\mathrm{Cov}(k) \quad (2)$$

Cov(k) quantifies the proportion of job descriptions in cluster k that contain at least one representative keyword from T_k. Higher coverage indicates that the keyword summary better reflects the cluster content.

As matching on a single keyword can be an insufficiently strict criterion, a stricter variant is also considered, specifically the thresholded coverage. For a threshold $\tau \geq 2$, thresholded coverage requires that a document match at least τ keywords from T_k:

$$\mathrm{Cov}_\tau(k) = \frac{1}{|C_k|}\sum_{d\in C_k} I\left(|W(d)\cap T_k| \geq \tau\right) \quad (3)$$

$$\mathrm{Cov}_\tau = \sum_{k=1}^{K}\frac{|C_k|}{N}\mathrm{Cov}_\tau(k) \quad (4)$$

This variant strengthens the representativeness requirement and reduces the possibility of inflated coverage caused by occasional generic term matches.

2) Outlier rate: The outlier rate quantifies the proportion of documents not explained by the topic summary under the same criterion:

$$\mathrm{Out}(k) = 1 - \mathrm{Cov}(k) \quad (5)$$

The corpus-level outlier rate is computed as:

$$\mathrm{Out} = 1 - \mathrm{Cov} \quad (6)$$

3) Overlap: Overlap measures discriminability by quantifying redundancy among topic summaries, using the Jaccard similarity between keyword sets T_i and T_j:

$$\mathrm{Jacc}(i,j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \quad (7)$$

The overall overlap is defined as the mean pairwise similarity:

$$\mathrm{Overlap} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le k} \mathrm{Jacc}(i,j) \quad (8)$$

The Jaccard similarity is a widely used metric in the analysis of set-based data [38]. It measures the proportion of shared elements relative to the union of sets, providing an interpretable quantification of similarity or redundancy. Its extensive application in computational linguistics and topic analysis establishes it as a reliable metric for evaluating the distinctiveness of topic keyword sets. Lower overlap indicates that different topics are characterized by more distinct keyword summaries, improving interpretability and reducing ambiguity in topic-to-competency translation. Because recruitment texts may share unavoidable common terms, overlap is interpreted jointly with coherence and coverage.

The proposed metrics jointly capture representativeness and discriminability aspects of topic outputs. It is important to clarify that these metrics do not aim to judge the linguistic aesthetics of competency labels, nor do they evaluate clustering quality in a supervised sense. Instead, the evaluation focuses on the structural validity of the evidence used for translation. At the task domain-level, the evaluation examines whether the Top-M topic summaries provide a stable, representative, and discriminative basis for inducing competencies. Coverage and outlier rate quantify how well the summary aligns with documents assigned to a topic in terms of representativeness and noise, and overlap indicates how distinguishable different topic summaries are by capturing redundancy across topics. At the competency level, the same validation logic is extended to the internal audit of the induced competencies. In this stage, the evaluation examines whether the translated competencies retain the dominant topic evidence and whether the decomposition remains internally coherent with limited redundancy. Collectively, these measures assess not only whether the topic-derived evidence is reliable enough to support competency induction, but also whether the resulting topic-to-competency translation remains traceable and interpretable under unlabeled conditions.

Topic summaries, metrics, and translation details are discussed in the Appendix.

III. Experimental Setup

This section describes the experimental setup used to evaluate the proposed translation framework. A baseline method is introduced to provide a conventional reference point, enabling a quantitative comparison that highlights the effectiveness and interpretability of the proposed approach. The baseline serves to contextualize the results and to demonstrate that observed improvements are attributable to the design of the proposed method rather than to trivial differences in data or partitioning.

A. Experimental Protocol

This section describes the evaluation protocol used to assess both the quality of topic evidence and the structural validity of the downstream competency translation. The protocol operates at two complementary levels. The first level evaluates topic summaries as the evidence basis for task domain induction, while the second level audits the resulting competencies derived from the proposed topic-to-competency translation procedure. This design is necessary because coherent topic summaries do not inherently guarantee that the resulting competency decomposition remains evidence-grounded, non-redundant, and traceable.

At the task domain-level, the protocol compares LDA topic discovery with a classical document clustering baseline under matched granularity and a unified interpretability interface. Comparability is enforced through four controls. First, all approaches operate on the same cleaned corpus and follow the same preprocessing and vocabulary construction procedure. Second, the number of topics or clusters is fixed to K across methods. Third, each topic or cluster is represented by a Top-M keyword set under a unified summary length. Fourth, all approaches produce document assignments and keyword summaries that are evaluated using the same metrics. Under this design, coverage, outlier rate, and overlap are used to assess whether the resulting summaries provide a representative and discriminative evidence interface for subsequent competency translation.

At the competency level, the evaluation shifts from topic summaries to the internal structure of the induced competency taxonomy. This level focuses on the transformation from topic evidence to a two-level representation consisting of task domains and competencies. Because the TF-IDF + K-Means baseline does not produce an explicit competency decomposition, a competency-level audit does not apply to the baseline and is therefore conducted only on the proposed LDA-based taxonomy. This audit examines whether the competencies preserve the dominant topic evidence and whether the decomposition remains structurally coherent within each task domain. Specifically, within-topic coverage is used to assess whether the highest-ranked topic keywords are retained in the competency structure; within-topic outliers are used to identify evidence loss during translation; and within-topic overlap among competency keyword subsets is used to examine internal redundancy. Taken together, the task domain-level comparison and the competency-level audit form a complementary validation protocol that evaluates not only the suitability of the topic evidence but also the interpretability and structural validity of the resulting task domain-competency hierarchy.

A traditional text clustering baseline is implemented using TF-IDF document representation followed by K-Means clustering. Based on the same preprocessed recruitment texts, each advertisement is represented as a TF-IDF vector and

partitioned into K clusters, where K corresponds to the number of topics used in the LDA model for direct comparison. TF-IDF is a standard weighting scheme in information retrieval and text mining. It assigns higher weights to terms that are frequent within a document but relatively less frequent across the corpus, which helps emphasize terms that are more discriminative for a given advertisement. In this study, TF-IDF vectors are constructed under a shared vocabulary derived from the same preprocessing pipeline. K-Means is a widely used unsupervised clustering method that partitions a set of vectors into K clusters by minimizing within-cluster dispersion. It iteratively assigns documents to cluster centers and updates the centers until the objective stabilizes. For each cluster, a fixed-length set of representative keywords is extracted from the cluster centroid. The centroid is obtained by aggregating the TF-IDF vectors of documents assigned to the cluster, and the Top-M terms with the highest centroid weights are selected as the cluster keywords. This centroid-based keyword extraction provides an interpretable summary for each cluster under the same Top-M interface used for topic summaries, enabling a consistent comparison across methods. This baseline therefore serves as a conventional reference for examining whether the evidence provides a more suitable basis for downstream competency translation than a classical clustering summary, rather than representing an alternative implementation of the proposed framework.

B. Implementation Details

Experiments are conducted on the robotics job advertisements corpus. The web crawling stage retrieved 6,802 advertisements, and 3,290 advertisements were retained after cleaning and deduplication. All methods operate on the same cleaned corpus to ensure that observed differences are attributable to modeling choices rather than data handling.

Latent competency topics were induced using the LatentDirichletAllocation class in scikit-learn. Before model training, the corpus was represented in document-term form, where min_df = 5 was used to remove extremely rare terms, and max_df = 0.9 was used to exclude overly common terms. The number of topics was specified as K, and the LDA hyperparameters were determined with reference to the expected concentration of the latent distributions and the practical requirements of model implementation. Specifically, the doc_topic_prior was set to $\alpha=50/K$, and the topic_word_prior was set to $\beta=0.01$. These settings were adopted as empirical parameter settings reported in prior LDA applications [39]. The model was trained with learning_method = 'batch', under which all documents in the corpus were used in each variational EM update, and random_state = 42 was fixed to ensure reproducibility. In addition, max_iter = 800 was specified as a sufficiently large upper bound on the number of training iterations.

For the TF-IDF representations used in the clustering baseline, vocabulary construction applied the same document frequency thresholds as in the LDA analysis, namely min_df = 5 and max_df = 0.9. The resulting shared vocabulary was used to construct TF-IDF vectors for clustering and to support comparable keyword summaries across methods. The K-Means baseline was configured with n_init = 11 and max_iter = 800. Multiple initializations were used to reduce sensitivity to local optima in sparse high-dimensional text representations. These settings were aligned with those used in the LDA analysis to facilitate controlled comparison across methods.

All experiments were set up on Windows 10 with Python 3.13.5. The implementation used Scikit-learn 1.6.1, NumPy 2.1.3, Pandas 2.2.3, Selenium 4.20.0, and Jieba 0.42.1.

IV. Results

A. Topic Discovery and Translation

The LDA model yields a set of competency-related topics, each summarized by a ranked list of top words derived from the topic-word distributions. To support subsequent competency translation, the number of topics is selected by jointly considering quantitative diagnostics and interpretability requirements. Following standard LDA practice, topic quality is evaluated using coherence and perplexity. Coherence quantifies the semantic consistency of the top words within each topic, based on word co-occurrence statistics computed from the tokenized corpus. Perplexity measures the model's probabilistic fit on holdout documents. Fig. 2 reports both criteria across a range of topic numbers.

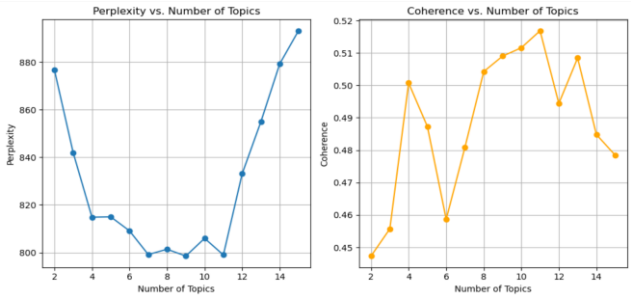


Fig. 2. Coherence and perplexity scores across topic numbers.

From a quantitative perspective, K=11 corresponded to the highest coherence score among the tested values, while perplexity remained within a relatively stable low range for $K \in [7,11]$. Taken together, these results indicate that the 11-topic model offered a suitable trade-off between topic interpretability and model fit among the candidate solutions. This suggests that the selected model not only fits the corpus adequately but also produces topic structures that are more coherent and more readily interpretable. From a modeling and application perspective, this level of granularity aligns with major robotics task domains and supports a two-level representation consisting of task domains and competencies. Fewer topics tend to merge heterogeneous skill signals, reducing discriminability at the task domain-level and complicating competency naming. More topics can fragment coherent domains into overly specific themes, increasing redundancy at the task domain-level and making the downstream competency decomposition less stable. The baseline comparison is carried out with the same partition count to ensure a fair evaluation across methods.

Based on this configuration, the topics are summarized by their top keywords and then translated into a task domain-competency taxonomy. Table I presents the resulting task

domains and competencies, including the top 30 keywords for each topic.

TABLE I. TOPIC-TO-COMPETENCY TRANSLATION RESULT

Task Domains	Competency	Top-30 Keywords
#1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis	1. Troubleshooting and Repair of Automated Equipment Malfunction	Equipment, Fault, Automation, Team, Issue, Mechanical, Technology, System, Collaboration, Fundamentals, Electrical, Production, AGV, Logging, Content, Data, Troubleshooting, Task, Problem Solving, Awareness, Professional Spirit, Skill, Optimization, Learning Ability, Testing, Teamwork, On-Site, Industrial, Sensor, Principle
	2. On-site Process Data Logging and Analysis	
	3. Teamwork and Problem Solving	
	4. Mechatronic System Principles Understanding and Technical Optimization	
	5. Industrial Sensor Testing and Equipment Maintenance	
#2 Mechanical Structure Design and Engineering Implementation	1. Mechanical Structure Design and Analysis	Design, Mechanical, Software, Process, Structure, Product, Mechanical Design, Analysis, Structural Design, Material, Selection, Automation, Machining, Technology, Equipment, CAD, Drive System, Issue, Production, Mechanism, Drawings, Principle, Work Experience, Project, Solution, Mechatronic, Optimization, Motor, Collaboration, Engineering
	2. Automation and Electromechanical Drive Systems Optimization	
	3. Material Selection and Manufacturing Process Planning	
	4. CAD Drafting and Design Documentation Output	
	5. Project Collaboration and Resolution of Product Technical Issues	
#3 Full Lifecycle Project Management and Operational Execution	1. Project Planning and Documentation Management	Product, Project, Technology, Team, Management, Industry, Customer, Quality, Analysis, Solution, Company, Process Flow, Issue, User, Market, Plan, Deployment, Project Management, Intelligent, Collaboration, Department, Documentation, Resources, Planning, Objective, Business, Design, Process, Solution, Automation
	2. Customer Requirements Analysis and Solution Design	
	3. Cross-Departmental Collaboration and Process Improvement Execution	
	4. Product Implementation Delivery and Technical Issue Resolution	
	5. Industry Resource Integration and Requirement Fulfillment	
#4 Robotics Core Module Selection Design and Testing	1. Circuit Design and Electronic Component Selection	Hardware, Design, Motor, Circuit, Electronics, Structure, Joint, Product, Testing, Module, Wiring Harness, Production, Team, Mechanical, Professional Spirit, Issue, Laboratory, Hands-On, Component, Attitude, Quality, Mindset, Power Supply, System, Circuit Design, Collaboration, Selection, Fundamentals, Management, Tooling
	2. Motor and Joint Module Integration with Wiring Harnesses	
	3. Hardware Prototyping, Testing, and Verification	
	4. Practical Engineering Skills and Team Collaboration	
#5 Electrical Design and System Integration for customized Automation	1. PLC Programming and System Software Debugging	Electrical, Automation, PLC, Project, Equipment, Programming, Design, Program, Customized, Software, Selection, Industrial, On-Site, Work Experience, Technology, System, Mechatronic, Electrical Design, Drawings, Vision, Robot Manipulator, Schematic Diagram, Communication, Principle, Wiring, Industry, Mitsubishi, Drafting, CAD, Signal
	2. CAD Drafting and Electrical Schematic Interpretation	
	3. On-site Wiring, Commissioning, and System Integration	
	4. Industrial Communication Configuration and Electromechanical Joint Debugging	
	5. Vision and Robot Manipulator Application in Custom Automation Projects	
#6 Embedded Hardware and Driver Software Development for Robotics	1. Embedded Linux Driver Programming and Development	Testing, System, Software, Driver, Automation, Design, Documentation, Programming, Sensor, Hardware, Embedded, Linux, Analysis, Technology, Issue, Computer, Project, Platform, Product, Motion, Functionality, Module, Principle, ROS, Python, Code, Electronics, Work Experience, Tooling, Motor
	2. Sensor and Motion Control Module Development	
	3. Test Analysis and Fault Localization	
	4. ROS System Design and Application Development	
	5. Project Documentation and Product Deliverable Handover	
#7 Robotics Algorithm Development and Deployment	1. Trajectory Planning and Control Algorithm Development	Algorithm, Planning, Model, Motion, Optimization, Data, Project, Technology, Python, Robotic Arm, Deep Learning, Framework, Robot, ROS, Kinematics, Deployment, Path, Domain, Real-World, Vision, Fundamentals, Control Algorithm, Slam, Environment, Scenario, AI, System, Mathematics, Training, Dynamics
	2. Deep Learning Model Training and Optimization	
	3. SLAM and Environmental Perception Algorithm Implementation	
	4. Algorithm Integration and Deployment in ROS Framework	
#8 Robotic System Performance Optimization and Reliability Testing	1. System Performance and Parameter Testing and Optimization	Optimization, Testing, Design, System, Technology, Performance, Issue, Product, Analysis, Solution, Data, Functionality, Documentation, Hardware, Stability, Reliability, Motion, Evaluation, Coordination, Module, Efficiency, Scenario, Team, Sensor, Planning, Process, Parameter, Review, Environment, Core
	2. Reliability-oriented Design and Stability Validation	
	3. Scenario-based Testing for Motion Modules and Sensors	
	4. Product Function Analysis and Test Documentation Preparation	
	5. Team Coordination and Efficiency-oriented Planning	
#9 Industrial Robot Process Programming and	1. Mainstream Industrial Robot Controllers Programming	Programming, Process, ABB, Industrial, Project, KUKA, Automation, FANUC, Work Experience, On-Site, System, Program, Automotive,
	2. Specific Process Application and Parameter Configuration	

Commissioning	3. Teamwork and Responsibility Awareness	YASKAWA, Brand, Offline, Sense Of Responsibility, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Teamwork, Laser, Body-In-White, Responsibility, Mechanical, Trajectory, Dispensing
#10 On-site Technical Support and After-sales Service	1. On-site Emergency Fault Diagnosis and Handling	On-Site, Customer, Project, Equipment, Technical Support, Issue, Product, Process, Production, Electrical, Company, After-Sales, Mechanical, Personnel, Solution, Technology, Process, Materials, Fault, Automation, Service, Organization, Guidance, Management, Optimization, Aspects, Tooling, File, Client, Department
	2. Customer Communication and Technical Service Provision	
	3. Materials Organization and After-sales Documentation Management	
	4. Project Service Delivery and Resource Coordination	
#11 Machine Vision Inspection System Solution Implementation	1. Image Processing Algorithm Programming	Vision, System, Machine Vision, Software, Project, Image Processing, Inspection, Hardware, Automation, Industrial, Design, Technology, Team, Algorithm, Selection, Customer, Light Source, Computer, Lens, Equipment, Analysis, Programming, Solution, OpenCV, Evaluation, Localization, Content, On-Site, VisionPro, Work Experience
	2. Vision Hardware Selection and Integration	
	3. Technical Solution Design and Evaluation	

B. Quantitative Comparison

Table II summarizes the task domain-level structural metrics for the LDA result and the TF-IDF plus K-Means clustering under the same protocol, including the same partition count $K = 11$ and the same summary length $M = 30$. The LDA topics achieve a coverage of 0.9739 with an outlier rate of 0.0261, indicating that the topic summaries remain broadly representative of the documents assigned to each topic. The baseline method yields a near-saturated weighted coverage of 0.9936 and an outlier rate of 0.0064, reflecting strong within-cluster lexical alignment between centroid-derived keywords and clustered documents. Although the LDA method attains a lower coverage than the baseline, the two coverage scores remain relatively close in magnitude, suggesting that both approaches provide broadly representative task domain-level evidence summaries under the same protocol.

A different pattern is observed in inter-cluster redundancy. The average Jaccard overlap of the keyword summaries is 0.1343 for LDA and 0.1906 for the baseline method. Lower overlap indicates that topic summaries are more distinguishable at the keyword level, which is beneficial for competency induction because redundancy across topics increases ambiguity in domain naming and reduces the separability of the induced competency domains. The higher overlap of the baseline method suggests that multiple clusters share a larger portion of their top keywords, which is consistent with centroid-based summaries favoring broadly applicable recruitment terms that recur across job families. In contrast to the gap in coverage, the difference in overlap is noticeably larger, indicating a more pronounced separation in the distinctiveness of the resulting task domain-level representations.

Taken together, the task domain-level results suggest that the baseline's high representativeness is accompanied by increased redundancy across clusters. The higher coverage achieved by the baseline is largely attributable to broadly used recruitment terms that appear frequently within clusters but also recur across multiple clusters, which inflates inter-cluster overlap and weakens topic distinctiveness. In contrast, LDA yields more distinctive topic summaries with lower overlap, resulting in a more balanced trade-off between representativeness and separability. Such a balance is particularly relevant for constructing a task domain-competency hierarchy that supports consistent competency induction across different robotics role categories and avoids redundancy.

TABLE II. COMPARATIVE TASK DOMAIN-LEVEL STRUCTURAL METRICS

Structural Metrics	Proposed Method (LDA)	Baseline (TF-IDF + K-Means)
Coverage ($\tau = 2$)	0.9739	0.9936
Outlier rate	0.0261	0.0064
Overlap	0.1343	0.1906

Beyond the task domain-level comparison, the proposed taxonomy is further examined through competency-level metrics that operate on the internal decomposition within each task domain. Unlike the task domain-level metrics, which evaluate whether the keyword summaries provide a representative and discriminative evidence basis for competency induction, the competency-level metrics evaluate whether the resulting competencies remain grounded in the original topic evidence after translation. Because the TF-IDF plus K-Means baseline does not generate an explicit competency structure, this audit is applied only to the proposed LDA-based taxonomy.

Table III reports the competency-level audit results. The results indicate that the proposed translation procedure retains a substantial proportion of the dominant task domain evidence at the competency level. The average within-topic coverage reaches 0.9667, suggesting that the highest-ranked topic signals are largely preserved in the translated competencies rather than being lost during decomposition. At the same time, the within-topic outlier remains 0.0333 on average, indicating that most evidence is incorporated into the competency structure rather than remaining outside the induced competencies. In addition, the average within-topic overlap among competency keyword subsets is 0.0246, showing that the decomposition does not merely restate the same semantic content under multiple labels but instead maintains relatively clear internal boundaries among competencies.

TABLE III. COMPETENCY-LEVEL AUDIT METRICS

Task Domain Number	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap
# 1	1.0000	0.0000	0.0000
# 2	0.9667	0.0333	0.0100
# 3	0.9333	0.0667	0.0000
# 4	0.9333	0.0667	0.0414
# 5	0.9333	0.0667	0.0268
# 6	0.9667	0.0333	0.0798
# 7	1.0000	0.0000	0.0781
# 8	0.9667	0.0333	0.0232
# 9	0.9667	0.0333	0.0000
# 10	1.0000	0.0000	0.0111
# 11	0.9667	0.0333	0.0000
Mean	**0.9667**	**0.0333**	**0.0246**

When interpreted jointly, the two levels of quantitative evidence support different but complementary claims. The task domain-level metrics show that the adopted LDA summaries provide a comparatively suitable evidence basis for competency induction in terms of representativeness and inter-cluster distinctiveness. The competency-level result further shows that, once translated, the competency-level structure preserves the dominant task domain evidence while limiting excessive within-topic repetition. Accordingly, the quantitative results do not establish an absolute notion of taxonomy correctness, but they do support the structural reliability, traceability, and interpretability of the proposed topic-to-competency translation under unlabeled conditions.

C. Qualitative Interpretation

Overall, the discovered topics cover major task domains frequently appearing in robotics recruitment texts, including industrial on-site operation and maintenance, mechanical design and engineering implementation, system integration and commissioning, embedded hardware and driver development, algorithm development and deployment, and performance validation and reliability testing. The keyword patterns provide a suitable evidence base for subsequent competency induction because they expose interpretable, task domain-level signals rather than only word frequency effects.

Based on the task domains, the associated competencies, and the Top-30 keyword evidence, the topics are organized into four groups. This organization provides a structured interpretation of the topic set and clarifies how topic evidence supports competency induction across development, deployment, and lifecycle operation.

The first group covers on-site operation and service and includes Domain #1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis, and Domain #10 On-site Technical Support and After-sales Service. These topics align with roles such as field operation and maintenance engineers, equipment maintenance engineers, and after-sales technical support engineers. The skill pattern is strongly contextual and workflow-oriented. High-frequency terms emphasize fault diagnosis, troubleshooting, and the maintenance of deployed equipment such as automated guided vehicle (AGV) systems. The topics also contain repeated signals related to after-sales service and data recording. Together, these signals indicate a closed loop of detection, restoration, documentation, and handover. This group is practically significant because it directly affects system availability in real deployments and influences recovery efficiency and downtime costs. The keyword evidence consistently highlights field constraints and operational accountability, indicating persistent demand for personnel who can restore deployed systems under real-world conditions while maintaining traceability of actions and outcomes.

The second group corresponds to engineering design and integration and spans a major engineering chain from design to delivery. It includes Domain #2 Mechanical Structure Design and Engineering Implementation, Domain #3 Full Lifecycle Project Management and Operational Execution, Domain #4 Robotics Core Module Selection Design and Testing, Domain #5 Electrical Design and System Integration for Customized Automation, Domain #9 Industrial Robot Process Programming and Commissioning, and Domain #11 Machine Vision Inspection System Solution Implementation. This group covers multiple role categories, including mechanical and hardware engineers, electrical and automation engineers, integration and commissioning engineers, project managers, solution engineers, and vision engineers. The skill characteristics reveal three connected capability streams. The first stream is mechanical and hardware engineering and is represented by Domain #2 and #4. It follows an engineering workflow centered on design, analysis, selection, prototyping, integration, and validation. The keyword evidence emphasizes mechanical design and analysis, materials and manufacturing planning, engineering drafting, circuit and component selection, and prototype testing. These capabilities define manufacturability, maintainability, and performance boundaries and constitute the engineering foundation of robotics products. The second stream is electrical design and system integration, and is represented by Domain #5 and #9. It emphasizes controller programming, on-site wiring and commissioning, industrial communication configuration, and electromechanical joint debugging, as well as process programming and parameter configuration for industrial robots. These competencies support the delivery of deployable automation solutions under industrial constraints. The third stream is solution delivery and project execution, and is represented by Domain #3 and #11. It highlights requirements analysis, solution design and evaluation, cross-department collaboration, process coordination, delivery execution, and documentation. These capabilities connect technical development to application deployment by transforming technical resources into deliverable solutions. Domain #11 further indicates combined demand for image processing programming, vision hardware selection, and engineering integration, reflecting role-specific requirements for machine vision solution implementation in industrial robotics.

The third group covers software and algorithm development and includes Domain #6 Embedded Hardware and Driver Software Development for Robotics and Domain #7 Robotics Algorithm Development and Deployment. These

topics align with robotics software engineers, embedded engineers, and algorithm engineers. Domain #6 emphasizes engineering-oriented development across embedded platforms, driver programming, sensor and motion module integration, testing and fault localization, and documentation-oriented delivery. Domain #7 focuses on motion planning and control, deep learning training and optimization, environmental perception, and deployment within robotics software frameworks. The combined signal indicates that industry demand extends beyond algorithm development and includes engineering integration and deployment. This group governs both the upper bound of system intelligence and the feasibility of transferring algorithms from development environments to production deployments with maintainable interfaces and stable performance.

The fourth group concerns system verification and reliability assurance and is represented by Domain #8, Robotic System Performance Optimization and Reliability Testing. It aligns with roles such as system test engineers, validation engineers, and reliability engineers. The topic evidence concentrates on reliability testing and performance testing. It also contains documentation and coordination signals that reflect process discipline in engineering validation. This group supports sustained stability and scalable deployment by enforcing measurable validation procedures and reducing deployment risks. It also forms a quality assurance bridge between development, integration, and operation.

In summary, the four groups jointly form a competency taxonomy that covers key role categories across development, delivery, and lifecycle operation. The topic-to-competency translation enhances interpretability by converting keyword summaries into competencies under task domains. Each competency remains traceable to a task domain-specific keyword subset, enabling interpretability of competency statements and supporting qualitative interpretation. This structured view complements the quantitative results by explaining how topic distinctiveness and representativeness support a competency taxonomy for robotics recruitment mining.

V. Conclusion and Future Prospects

A. Contributions and Practical Implications

This study contributes to the literature on recruitment text mining and competency induction by proposing a data-driven approach for translating robotics recruitment texts into a structured competency taxonomy. The main outcome is a two-level competency taxonomy comprising 11 task domains and 48 competencies, grounded in topic keyword evidence and supported by a translation procedure and a traceable validation logic.

This study identifies latent competency domains and skill patterns embedded in robotics recruitment texts and organizes them into a coherent structure. The resulting taxonomy provides an empirical representation of employer competency demand across major robotics role families, including development, integration, deployment, and operation. Rather than remaining a list of isolated skills, the identified taxonomy is organized into a hierarchical structure that makes the relationships among task domains and competencies more explicit. Building on the robotics talent competency literacy framework analyzed in [13], the present study not only supports its main perspective but also further extends and enriches the skill requirements identified for the robotics domain, thereby producing a more detailed and fine-grained competency inventory. This richer and more structured competency taxonomy expands existing research on job-related competencies in the robotics field and provides a more actionable basis for occupational analysis and curriculum design.

Moreover, the study extracts latent topics from recruitment texts through LDA and then applies a translation procedure to convert unsupervised topic evidence into task domains and competencies under explicit requirements of keyword coverage and redundancy control. With explicit translation constraints and metrics, this approach further reduces the problem of naming subjectivity discussed in [25]. It also enriches the methodological pathway discussed in [21][22][23] for mining competencies through LDA and related models. The contribution of the study therefore lies not only in generating topic summaries, but also in providing an approach from topic keyword evidence to a competency taxonomy that can support both labor market analysis and educational planning.

The study also has important practical implications for labor market analysis, vocational education, and workforce development. First, the resulting taxonomy provides a traceable competency artifact that can support structured analysis of robotics labor demand, including role profiling, demand monitoring, and the identification of major competency clusters in industry recruitment. Because the taxonomy is derived from employer recruitment texts rather than predefined occupational templates, it reflects current market competency signals and can therefore complement more static competency frameworks developed through expert judgment.

Second, the study has reference value for talent cultivation in vocational education and for the development of teaching reform. Because the taxonomy is grounded in observed employer demand, it can provide an empirical reference for curriculum alignment, competency module refinement, training pathway design, and the identification of emerging skill combinations that may not yet be fully reflected in conventional educational programs. In robotics education, where the integration of mechanics, electronics, control, software, and application deployment is especially important, a competency structure grounded in observed demand can help connect educational planning more closely to industrial requirements.

Overall, the study contributes not only a robotics competency taxonomy but also a methodological perspective on how recruitment text evidence can be translated into an interpretable competency structure. Its value lies less in claiming superiority of a single topic model and more in demonstrating how unsupervised topic evidence can be organized, interpreted, and evaluated as a competency representation under realistic unlabeled conditions.

B. Limitations and Future Work

The scope of the present study is subject to the following limitations. First, the recruitment corpus is collected from a bounded set of online sources within the Chinese labor market, with job advertisements sampled from several representative major cities. This scope implies constraints on external generalizability. The resulting competency taxonomy should therefore not be directly transferred to overseas regions with different industrial structures, regulatory environments, or technological development stages. In addition, the corpus is gathered within a limited sampling period. In a rapidly evolving domain such as robotics, the observed skill demand signals primarily reflect the workforce requirements at the time of collection, and the competency profile is expected to shift as technologies mature and deployment patterns change.

Second, the evaluation emphasizes internal interpretability and structural quality using coherence and other metrics derived from topic and keyword evidence, including coverage, outlier rate, and overlap. This design matches the study setting in which labels for topic quality and competency taxonomy correctness are typically unavailable in recruitment text, and it enables comparison across methods under a unified evidence interface. The current results therefore provide internal structural support for the induced competency taxonomy, particularly in terms of representativeness, distinctiveness, and traceability. However, the present validation does not yet extend to external substantive assessment by domain experts. Additional external evidence, such as expert review, could further strengthen the interpretive claims and broaden the validation perspective.

Future research can extend the present study along two directions. First, external validity and longitudinal analysis can be strengthened through a structured cross-cultural validation framework. This includes expanding data collection to multiple countries or regions and conducting comparisons across markets with differing industrial contexts. Longitudinal designs should be prioritized to track how competency demand evolves, enabling the separation of stable competency cores from time-sensitive requirements and emerging skill trends. Second, external evidence and application validation can be incorporated as a complementary verification layer beyond internal interpretability metrics. These extensions may incorporate expert review supported by consensus analysis, which can further reduce the influence of subjective naming choices beyond the translation constraints imposed by the translation procedure, and validation through downstream tasks that require structured competency representations.

ACKNOWLEDGMENT

This research was supported by Jiangsu Province Higher Education Teaching Reform Research Project (Grant No. 2025JGYB674), Teaching Reform Project of Suzhou Vocational University (Grant No. SZDJG-24006), and Special Project of Suzhou Polytechnic University (Grant No. GJSZX-260110).

REFERENCES

[1] Alhloul A, Kiss E. Industry 4.0 as a Challenge for the Skills and Competencies of the Labor Force: A Bibliometric Review and a Survey[J]. Sci, 2022, 4(3): 34.

[2] Kipper L M, Iepsen S, Dal Forno A J, et al. Scientific mapping to identify competencies required by industry 4.0[J]. Technology in Society, 2021, 64: 101454.

[3] Sheriff N, Sevukan R. Discovering research data management trends from job advertisements using a text-mining approach[J]. Journal of Information Science, 2023: 01655515231193845.

[4] Wowczko I A. Skills and vacancy analysis with data mining techniques[C]//Informatics. MDPI, 2015, 2(4): 31-49.

[5] Hernandez-de-Menendez M, Morales-Menendez R, Escobar C A, et al. Competencies for industry 4.0[J]. International Journal on Interactive Design and Manufacturing (IJIDeM), 2020, 14(4): 1511-1524.

[6] Maisiri W, Darwish H, Van Dyk L. An investigation of industry 4.0 skills requirements[J]. South African Journal of Industrial Engineering, 2019, 30(3): 90-105.

[7] Nasir S A M, Wan Yaacob W F, Wan Aziz W A H. Analysing online vacancy and skills demand using text mining[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1496(1): 012011.

[8] Tzimas G, Zotos N, Mourelatos E, et al. From data to insight: Transforming online job postings into labor-market intelligence[J]. Information, 2024, 15(8): 496.

[9] MacKay M, Ford C, Grant L E, et al. Developing competencies in public health: a scoping review of the literature on developing competency frameworks and student and workforce development[J]. Frontiers in public health, 2024, 12: 1332412.

[10] Kiesler N. Towards a competence model for the novice programmer using bloom's revised taxonomy-an empirical approach[C]//Proceedings of the 2020 acm conference on innovation and technology in computer science education. 2020: 459-465.

[11] Irvine J. Taxonomies in education: Overview, comparison, and future directions[J]. Journal of Education and Development, 2021, 5(2): 1.

[12] Wang L, Ning Y, Li Z. Exploration and Practice of the Blended Teaching Mode of Industrial Robot Courses Based on the Background of Artificial Intelligence[C]//Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference. 2025: 802-806.

[13] Yu-Shen F, Shi-Jie W, Run-Jiao H, et al. Analysis of Core Competency Literacy of Innovative Technology Talents in Intelligent Robot Industry[C]//2020 5th International Conference on Humanities Science and Society Development (ICHSSD 2020). Atlantis Press, 2020: 106-110.

[14] Olukanni E, Akanmu A, Jebelli H. Industry perception of competencies for human—robot collaboration in the construction industry: A Delphi study[J]. Frontiers of Engineering Management, 2025, 12(4): 854-879.

[15] Do H D, Tsai K T, Wen J M, et al. Hard skill gap between university education and the robotic industry[J]. Journal of Computer Information Systems, 2023, 63(1): 24-36.

[16] Bademosi F M, Issa R R A. Essential knowledge, skills, and abilities required for talent cultivation in construction automation and robotics[M]//Automation and robotics in the architecture, engineering, and construction industry. Cham: Springer International Publishing, 2022: 31-57.

[17] Salotti J M, Suhir E. Collaborative Robotics: Application of Delphi Method[J]. Journal of Field Robotics, 2025, 42(5): 1799-1807.

[18] Gavrilescu M, Leon F, Minea A A. Techniques for transversal skill classification and relevant keyword extraction from job advertisements[J]. Information, 2025, 16(3): 167.

[19] Chen Y, Pan R. Research on data analysis and visualization of recruitment positions based on text mining[J]. Advances in Multimedia, 2022, 2022(1): 9047202.

[20] Colombo E, Mercorio F, Mezzanzanica M. Applying machine learning tools on web vacancies for labour market and skill analysis[J]. Terminator or the Jetsons, 2018.

[21] Gurcan F, Soylu A, Khan A Q. Towards a sustainable workforce in big data analytics: Skill requirements analysis from online job postings using neural topic modeling[J]. Sustainability, 2025, 17(20): 9293.

[22] Kang P S, Enstroem R, Bhawna B, et al. A text mining study of competencies in modern supply chain management with skillset mapping[J]. Supply Chain Analytics, 2025, 10: 100117.

[23] Gurcan F, Cagiltay N E. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling[J]. IEEE access, 2019, 7: 82541-82552.

[24] Djunaidi K, Kusuma D T, Ningrum R F, et al. Big Data Analytics of Knowledge and Skill Sets for Web Development Using Latent Dirichlet Allocation and Clustering Analysis[J]. International Journal of Advanced Computer Science & Applications, 2025, 16(1).

[25] Zhou Y, Tao L, Xue Z, et al. From Job Postings to Vocational Education Standards: Mapping Competency Requirements for NEV Sales and Livestreaming Hosts[J]. World Electric Vehicle Journal, 2026, 17(3): 162.

[26] Madsen A, Reddy S, Chandar S. Post-hoc interpretability for neural nlp: A survey[J]. ACM Computing Surveys, 2022, 55(8): 1-42.

[27] Soare E. Perspectives on designing the competence based curriculum[J]. Procedia-Social and Behavioral Sciences, 2015, 180: 972-977.

[28] McClelland D C. Identifying competencies with behavioral-event interviews[J]. Psychological science, 1998, 9(5): 331-339.

[29] Hoyle A, Goel P, Hian-Cheong A, et al. Is automated topic model evaluation broken? the incoherence of coherence[J]. Advances in neural information processing systems, 2021, 34: 2018-2033.

[30] Kim H, Park H, Song M. Developing a topic-driven method for interdisciplinarity analysis[J]. Journal of informetrics, 2022, 16(2): 101255.

[31] Fang F, Zhou Y. A study on recruitment of data analyst based on text mining and visualization technology[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1952(4): 042017.

[32] Zeng X, Chu S, Chen X. China's labor market demand in the shadow of COVID-19: Evidence from an online job board[J]. Journal of Asian Economics, 2025, 97: 101885.

[33] Gojare S, Joshi R, Gaigaware D. Analysis and design of selenium webdriver automation testing framework[J]. Procedia Computer Science, 2015, 50: 341-346.

[34] Thooriqoh H A, Annisa T N, Yuhana U L. Selenium framework for web automation testing: A systematic literature review[J]. JUTI: Jurnal Ilmiah Teknologi Informasi, 2021: 65-76.

[35] Cao S. New word detection algorithm combining correlation confidence and jieba word segmentation[J]. Comput Syst Appl, 2020, 29: 144-51.

[36] Na D, Xu C. Automatically generation and evaluation of stop words list for Chinese patents[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2015, 13(4): 1414-1421.

[37] Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia tools and applications, 2019, 78(11): 15169-15211.

[38] Bag S, Kumar S K, Tiwari M K. An efficient recommendation generation using relevant Jaccard similarity[J]. Information Sciences, 2019, 483: 53-64.

[39] Heinrich G. Parameter estimation for text analysis[R]. Darmstadt, Germany: Technical report, 2005.

APPENDIX. TOPIC SUMMARIES, METRICS, AND TRANSLATION DETAILS

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#1 Industrial On-site Equipment Maintenance and Malfunction Diagnosis	1. Troubleshooting and Repair of Automated Equipment Malfunction	Equipment, Fault, Automation, Team, Issue, Mechanical, Technology, System, Collaboration, Fundamentals, Electrical, Production, AGV, Logging, Content, Data, Troubleshooting, Task, Problem Solving, Awareness, Professional Spirit, Skill, Optimization, Learning Ability, Testing, Teamwork, On-Site, Industrial, Sensor, Principle	1.0000	0.0000	0.0000	Equipment, Fault, Automation, Electrical, AGV, Troubleshooting
	2. On-site Process Data Logging and Analysis					Production, Logging, Content, Data, On-Site
	3. Teamwork and Problem Solving					Team, Collaboration, Problem Solving, Awareness, Professional Spirit, Learning Ability, Teamwork, Task
	4. Mechatronic System Principles Understanding and Technical Optimization					Issue, Mechanical, Technology, System, Skill, Optimization, Principle
	5. Industrial Sensor Testing and Equipment Maintenance					Testing, Industrial, Sensor,Maintenance
#2 Mechanical Structure Design and Engineering Implementation	1. Mechanical Structure Design and Analysis	Design, Mechanical, Software, Process, Structure, Product, Mechanical Design, Analysis, Structural Design, Material, Selection, Automation, Machining, Technology, Equipment, CAD, Drive System, Issue, Production, Mechanism, Drawings, Principle, Work Experience, Project, Solution, Mechatronic, Optimization, Motor, Collaboration, Engineering	0.9667	0.0333	0.0100	Design, Mechanical, Structure, Mechanical Design, Analysis, Structural Design, Principle
	2. Automation and Electromechanical Drive Systems Optimization					Automation, Equipment, Drive System, Mechanism, Mechatronic, Optimization, Motor
	3. Material Selection and Manufacturing Process Planning					Process, Material, Selection, Machining, Production
	4. CAD Drafting and Design Documentation Output					Software,CAD, Drawings, Solution, Engineering
	5. Project Collaboration and Resolution of Product Technical Issues					Product, Technology, Issue, Principle, Project, Collaboration
#3 Full Lifecycle	1. Project Planning and Documentation Management	Product, Project, Technology, Team,	0.9333	0.0667	0.0000	Project, Management, Plan, Project

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
Project Management and Operational Execution		Management, Industry, Customer, Quality, Analysis, Solution, Company, Process Flow, Issue, User, Market, Plan, Deployment, Project Management, Intelligent, Collaboration, Department, Documentation, Resources, Planning, Objective, Business, Design, Process, Solution, Automation				Management, Documentation, Planning, Process
	2. Customer Requirements Analysis and Solution Design					Customer, Analysis, Solution, Objective, Design, Solution
	3. Cross-Departmental Collaboration and Process Improvement Execution					Team, Company, Process Flow, Collaboration, Department, Business
	4. Product Implementation Delivery and Technical Issue Resolution					Product, Technology, Quality, Issue, Deployment
	5. Industry Resource Integration and Requirement Fulfillment					Industry, User, Market, Resources
#4 Robotics Core Module Selection Design and Testing	1. Circuit Design and Electronic Component Selection	Hardware, Design, Motor, Circuit, Electronics, Structure, Joint, Product, Testing, Module, Wiring Harness, Production, Team, Mechanical, Professional Spirit, Issue, Laboratory, Hands-On, Component, Attitude, Quality, Mindset, Power Supply, System, Circuit Design, Collaboration, Selection, Fundamentals, Management, Tooling	0.9333	0.0667	0.0414	Design, Circuit, Electronics, Product, Component, System, Circuit Design, Selection
	2. Motor and Joint Module Integration with Wiring Harnesses					Motor, Structure, Joint, Module, Wiring Harness, Mechanical, Power Supply, System
	3. Hardware Prototyping, Testing, and Verification					Hardware, Product, Testing, Mechanical
	4. Practical Engineering Skills and Team Collaboration					Team, Professional Spirit, Issue, Laboratory, Hands-On, Attitude, Mindset, Collaboration, Fundamentals, Management, Tooling
#5 Electrical Design and System Integration for customized Automation	1. PLC Programming and System Software Debugging	Electrical, Automation, PLC, Project, Equipment, Programming, Design, Program, Customized, Software, Selection, Industrial, On-Site, Work Experience, Technology, System, Mechatronic, Electrical Design, Drawings, Vision, Robot Manipulator, Schematic Diagram, Communication, Principle, Wiring, Industry, Mitsubishi, Drafting, CAD, Signal	0.9333	0.0667	0.0268	PLC, Programming, Program, System, Mitsubishi
	2. CAD Drafting and Electrical Schematic Interpretation					Electrical, Design, Software, Electrical Design, Drawings, Schematic Diagram, Principle, Wiring, Drafting, CAD
	3. On-site Wiring, Commissioning, and System Integration					Equipment, Industrial, On-Site, System
	4. Industrial Communication Configuration and Electromechanical Joint Debugging					Mechatronic, Industrial, Communication, Signal
	5. Vision and Robot Manipulator Application in Custom Automation Projects					Automation, Project, Customized, Selection, Technology, Vision, Robot Manipulator
#6 Embedded Hardware and Driver Software Development for Robotics	1. Embedded Linux Driver Programming and Development	Testing, System, Software, Driver, Automation, Design, Documentation, Programming, Sensor, Hardware, Embedded, Linux, Analysis, Technology, Issue, Computer, Project, Platform, Product, Motion, Functionality, Module, Principle, ROS, Python, Code, Electronics, Work Experience, Tooling, Motor	0.9667	0.0333	0.0798	Software, Driver, Programming, Hardware, Embedded, Linux, Technology, Computer, Principle, Code, Tooling
	2. Sensor and Motion Control Module Development					Software, Sensor, Technology, Motion, Functionality, Module, Principle, Code, Electronics, Tooling, Motor
	3. Test Analysis and Fault Localization					Testing, Automation, Analysis, Issue
	4. ROS System Design and Application Development					System, Software, Design, Technology, Platform, Principle, ROS, Python, Code, Tooling
	5. Project Documentation and Product Deliverable Handover					Documentation, Project, Product

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#7 Robotics Algorithm Development and Deployment	1. Trajectory Planning and Control Algorithm Development	Algorithm, Planning, Model, Motion, Optimization, Data, Project, Technology, Python, Robotic Arm, Deep Learning, Framework, Robot, ROS, Kinematics, Deployment, Path, Domain, Real-World, Vision, Fundamentals, Control Algorithm, Slam, Environment, Scenario, AI, System, Mathematics, Training, Dynamics	1.0000	0.0000	0.0781	Algorithm, Planning, Motion, Python, Robotic Arm, Kinematics, Path, Control Algorithm, Dynamics
	2. Deep Learning Model Training and Optimization					Model, Optimization, Data, Deep Learning, Framework, Robot, AI, Mathematics, Training, Domain
	3. SLAM and Environmental Perception Algorithm Implementation					Algorithm, Project, Python, Real-World, Vision, SLAM, Environment, Scenario
	4. Algorithm Integration and Deployment in ROS Framework					Algorithm, Technology, Python, Framework, ROS, Deployment, Fundamentals, System
#8 Robotic System Performance Optimization and Reliability Testing	1. System Performance and Parameter Testing and Optimization	Optimization, Testing, Design, System, Technology, Performance, Issue, Product, Analysis, Solution, Data, Functionality, Documentation, Hardware, Stability, Reliability, Motion, Evaluation, Coordination, Module, Efficiency, Scenario, Team, Sensor, Planning, Process, Parameter, Review, Environment, Core	0.9667	0.0333	0.0232	Optimization, Testing, System, Performance, Data, Hardware, Parameter
	2. Reliability-oriented Design and Stability Validation					Design, Stability, Reliability, Technology, Issue
	3. Scenario-based Testing for Motion Modules and Sensors					Testing, Motion, Module, Scenario, Sensor, Environment
	4. Product Function Analysis and Test Documentation Preparation					Testing, Product, Analysis, Solution, Functionality, Documentation, Evaluation, Review
	5. Team Coordination and Efficiency-oriented Planning					Coordination, Efficiency, Team, Planning, Process
#9 Industrial Robot Process Programming and Commissioning	1. Mainstream Industrial Robot Controllers Programming	Programming, Process, ABB, Industrial, Project, KUKA, Automation, FANUC, Work Experience, On-Site, System, Program, Automotive, YASKAWA, Brand, Offline, Sense Of Responsibility, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Teamwork, Laser, Body-In-White, Responsibility, Mechanical, Trajectory, Dispensing	0.9667	0.0333	0.0000	Programming, ABB, Industrial, KUKA, FANUC, Program, YASKAWA, Brand, Mechanical, Trajectory
	2. Specific Process Application and Parameter Configuration					Process, Project, Automation, On-Site, System, Automotive, Offline, Arc Welding, Mechatronic, Parameter, Configuration, Optimization, Spot Welding, Laser, Body-In-White, Dispensing
	3. Teamwork and Responsibility Awareness					Sense Of Responsibility, Teamwork, Responsibility
#10 On-site Technical Support and After-sales Service	1. On-site Emergency Fault Diagnosis and Handling	On-Site, Customer, Project, Equipment, Technical Support, Issue, Product, Process, Production, Electrical, Company, After-Sales, Mechanical, Personnel, Solution, Technology, Process, Materials, Fault, Automation, Service, Organization, Guidance, Management, Optimization, Aspects, Tooling, File, Client, Department	1.0000	0.0000	0.0111	On-Site, Equipment, Technical Support, Issue, Process, Electrical, Mechanical, Fault, Automation
	2. Customer Communication and Technical Service Provision					Customer, Product, Technology, Service, Guidance, Tooling, Client, Optimization
	3. Materials Organization and After-sales Documentation Management					Production, After-Sales, Materials, Organization, Management, File
	4. Project Service Delivery and Resource Coordination					Project, Company, Personnel, Solution, Process, Resource, Department

Task Domains	Competency	Top-30 Keywords	Within-topic Coverage	Within-topic Outlier	Within-topic Overlap	Assigned Keywords
#11 Machine Vision Inspection System Solution Implementation	1. Image Processing Algorithm Programming	Vision, System, Machine Vision, Software, Project, Image Processing, Inspection, Hardware, Automation, Industrial, Design, Technology, Team, Algorithm, Selection, Customer, Light Source, Computer, Lens, Equipment, Analysis, Programming, Solution, OpenCV, Evaluation, Localization, Content, On-Site, VisionPro, Work Experience	0.9667	0.0333	0.0000	Machine Vision, Software, Image Processing, Algorithm, Computer, Analysis, Programming, OpenCV, Localization, VisionPro
	2. Vision Hardware Selection and Integration					Vision, System, Inspection, Hardware, Industrial, Selection, Light Source, Lens, Equipment
	3. Technical Solution Design and Evaluation					Project, Automation, Design, Technology, Team, Customer, Solution, Evaluation, Content, On-Site

Enhancing User Experience in Virtual Reality for Children with Autism: A Significant Review

Nur Aleesya Mohd Asri[1], Normala Rahim[2*], Norsuhaily Abu Bakar[3],
Wan Rizhan[4], Ismahafezi Ismail[5], Nur Saadah Mohd Shapri[6], Sarah Farhana Juhari[7]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut, Terengganu, Malaysia[1, 2, 4, 5, 6]
Faculty of Applied Social Science, Universiti Sultan Zainal Abidin, Gong Badak, Terengganu, Malaysia[3]
Malaysia Digital Economy Corporation (MDEC), Cyberjaya, Selangor[7]

***Abstract*—Virtual reality (VR) has become a promising technology to assist children with autism spectrum disorder (ASD especially in the areas of improving user experience with immersive, adaptive, and interactive environments. Nevertheless, current research is still disjointed in areas with very little synthesis on the effectiveness of VR in enhancing user experience and supporting the various cognitive, behavioural, and sensory requirements of autistic children. This study seeks to thoroughly review and synthesize current developments in VR-based solutions to determine the main methods, assessment plans, and design implications that can enhance user experience in this population. The systematic literature review was based on the PRISMA framework, where an advanced search strategy was implemented in Scopus and IEEE databases with keywords related to autism, user experience, and virtual reality. The identification phase yielded 939 records, followed by a rigorous screening and eligibility process resulting in a final dataset of 29 primary studies published between 2020 and 2025. The results were divided into three broad themes, namely: 1) VR-Based Intervention and Skill Development for Autism which shows how social, cognitive and daily living skills can be improved with the help of an immersive training; 2) Assessment, Evaluation and Multimodal Analytics in VR for Autism which reveals how objective evaluation and personalisation can be achieved with the assistance of behavioural and physiological and AI driven data which can improve the evaluation and personalization; and 3) Design Framework which includes VR system for Autism that focuses on users, ethical consideration and an inclusive system development. Altogether, the review indicates that VR demonstrates substantial potential to improve the user experience of children with autism, especially when it is facilitated through adaptive, data-driven, and inclusive design strategies. However, issues of scalability, long-term efficacy, and standardisation persist, meaning that future studies are required to create strong and user-friendly VR frameworks in this population.**

***Keywords*—Virtual Reality (VR); Autism Spectrum Disorder (ASD); User Experience (UX); multimodal analytics; inclusive design**

I. Introduction

Virtual Reality (VR) is a technology that simulates lifelike and immersive experiences within a digitally created environment. Objects within VR tend to behave much as they do in the physical world, allowing users to engage with them in a manner that follows natural laws of physics [1], [2]. Hence, it has emerged as a groundbreaking technology in therapeutic and educational interventions in children with autism spectrum disorder (ASD) and offers immersive and interactive environments that can be tailored to particular needs. The role of user experience (UX) in this context is immeasurable, and the success of VR-based interventions depends on how well they can address the specific sensory, cognitive, and social aspects of autistic children. The increasing prevalence of ASD, together with limitations in traditional therapeutic approaches, has led to the rise of interest in applying VR to support social communication, emotional regulation, and adaptive behavior. However, the possibility of VR is limited by the difficulty of developing a digital space that is appealing and favorable to children with different sensory backgrounds and interaction. The review presents the current body of research on the usefulness of UX enhancement in VR among autistic children, focusing on the connections between theoretical assumptions and technological innovation and the practical implementation of this methodology in a clinical setting and in the educational process. [3], [4], [5].

Autistic children are known to have abnormal sensory processing behavior where they experience hyper or hyposensitivity to visual, auditory, and tactile stimuli [6]. Such sensory differences can greatly influence their tolerance and acclimatization to VR environments, where uncontrolled or excessive stimulation can result in distress, anxiety or disengagement [7]. Repeated studies have demonstrated that personalised sensory modulation, such as manipulation of the intensity, type and timing of sensory input may be used to alleviate sensory overload and facilitate a positive interaction. [8] Indicatively, VR systems as screens, delivering controlled multimodal stimuli have been shown to be tolerable to autistic adolescents and have been shown to have behavioral responses, which are correlated to individual sensory profiles [9], [10], [11]. Moreover, the emotional factors play a decisive role in the sensory responsiveness, and tactile hyper- or hypo-responsivity is typically coupled with anxiety or disparities in emotional processing in comparison with the mere perceptual incompetence [12], [13]. The VR design that is effective in this population is therefore not only technically precise in terms of sensory delivery but also ought to be sensitive to the emotional and behavioral context in which senses are utilized. [1]

The presence of engagement barriers and preferences for interaction further complicates the design of VR for autistic children. ASD children prefer natural, effective modalities of interaction, such as virtual hand controls or gesture interfaces, to formal game controllers, which can lessen cognitive load and

*Corresponding author.

improve task performance [14], [15]. Engagement may, however, be destroyed by unpredictability, multifaceted social signals and the danger of sensory bombardment necessitating predictable and adaptable contexts. VR worlds can be modified on-the-fly, depending upon physiological and behavioral feedback such as eye tracking and heart rate to make interventions more personalized and enable long-term engagement [16], [17]. Autistic children, educators and therapists have also been mentioned as critical in the co-design strategies, where they are involved in the development process and can create inclusive, comfortable and motivating VR experiences [18], [19]. Such participatory practices not only increase usability but also tackle ethical issues by making sure that interventions are respectful, relevant, and in line with the lived experiences of autistic users [20], [21].

VR technology in ASD is varied and includes a variety of hardware platforms such as head-mounted displays (HMDs) and desktop and projector-based systems, each with its own strengths and weaknesses. The interaction provided by an immersive VR using HMDs tends to be more engaging and spatial and can fit the visuospatial advantages of most autistic children [22], [23]. They may, however, also increase sensory sensitivities and cause cybersickness, especially in individuals with a high perceptual sensitivity, requiring specific desensitization measures and session control [24]. Less immersive or screen-based VR can provide a more accessible option to younger or more sensitive users but could potentially come at the cost of decreased immersion and engagement [24], [25]. On all platforms, customizable sensory parameters, spatialized audio, haptics and real-time adaptation frameworks are critical to lessening sensory overburden and maximizing therapeutic value [11], [17]. The evidence base of the VR interventions in ASD is strong and demonstrates the short-term outcomes of the interventions in improving social communication, cognitive and adaptive behaviors, yet there are still methodological limitations such as small sample sizes, lack of diversity and follow-ups [26], [27], [28].

Nevertheless, despite the potential of VR, significant research and technological gaps remain. The rigor of the methodology is often subpar, and the majority of the studies rely on small and homogeneous samples and short-term outcome measurements, limiting the possibility of generalizing and scaling up the findings [27], [28]. VR used together with established types of therapy, such as cognitive behavioral therapy, has been shown to have a synergistic effect on enhancing social, cognitive and emotional outcomes, although more systematic studies are still required to identify best practice and long-term effectiveness [29], [30]. The issue of the sensory overload, cybersickness, privacy, consent and the risk of interventions trying to normalize autistic characteristics rather than letting people be themselves are of paramount importance [20], [31]. The participatory design model and responsible research and innovation models are also increasingly mentioned as a requirement to ensure that VR technologies are created and implemented in a way that is both safe, inclusive and responsive to the needs of autistic children [21], [31]. Price, infrastructure, teacher education and individualized content barriers to greater implementation in education and clinical practice should also be surmounted through collaborative, user-friendly innovation and policy facilitation [32], [33].

This systematic review aims to address these gaps and is structured into several sections to provide a systematic understanding of virtual reality applications for children with autism. The next section of this study presents the research methodology based on the PRISMA framework, including the search strategy, screening criteria, eligibility process, and quality assessment procedures. Mixed Methods Appraisal Tool (MMAT) was conducted to evaluate the methodological quality of included studies and distinguish between high-quality and low-quality evidence. This is followed by the results and findings section, which discusses publication trends and thematic analysis derived from the selected studies. The thematic findings are categorized into three major themes: VR-based intervention and skill development, multimodal assessment and analytics, and inclusive design frameworks for autism-related VR systems. Finally, the discussion and conclusion section summarizes the key findings, highlight the practical and theoretical implications of the review, identifies current limitations, and proposes future research directions for enhancing user experience in VR environments for children with autism.

Overall, enhancing user experience in virtual reality for children with autism requires an integrative approach that bridges technical innovation with a comprehensive understanding of autistic users' sensory, cognitive, and emotional needs. In the literature, having personalized, adaptive VR spaces to embrace individual sensory profiles, leverage natural interaction patterns and co-created with people and stakeholders is emphasized to be relevant, both ethically and practically. Despite the gigantic potential of immersive VR technologies to improve social, cognitive and emotional outcomes, the utility of the latter should be justified by a serious methodological analysis, their confirmation by a certain time and developing the means of their implementation, which should be scalable and accessible. Further research is required to be more specific to the application of a wide range of population groups of participants, employing powerful longitudinal designs and integrating machine learning and real-time feedback systems that can help to further customize and optimize VR experiences. Finally, the fulfilment of the promise of VR to help autistic children will require long-term interdisciplinary cooperation, ethical responsibility and a focus on participatory and user-centered design [34], [35].

II. Materials and Methods

A. Identification

According to the PRISMA framework, the identification phase is a fundamental step to the comprehensiveness and transparency of a systematic literature review (SLR). Table I demonstrates that a systematic and thorough search approach was carried out in two major and complementary academic sources, including Scopus and IEEE Xplore, on a mixture of keywords regarding autism, user experience, and virtual reality. The initial search led to a total of 939 records, including 757 in Scopus and 182 in IEEE.

Such a distribution is suggestive of Scopus as a more comprehensive source of disciplinary coverage because it indexes a wide range of journals not only in the health sciences but also in psychology and the social sciences, and also in studies in technology. On the other hand, the IEEE Xplore is more technical and engineering-based and therefore offers a smaller scope of literature. The fact that there were a comparatively higher number of records found in Scopus indicates that the research on virtual reality applications in children with autism is not restricted to technical research but is extensively studied within the context of interdisciplinary research, including human-computer interaction, rehabilitation and educational technology.

TABLE I. SEARCH STRING

Scopus	TITLE-ABS-KEY (("autism" OR "autistic disorder" OR "autistic spectrum disorder" OR "ASD") AND ("virtual reality" OR "VR" OR "immersive technology" OR "simulated environment") AND ("user experience" OR "UX" OR "user-centered design" OR "interaction" OR "engagement")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBYEAR , "2020") OR LIMIT-TO (PUBYEAR , "2022") OR LIMIT-TO (PUBYEAR , "2023") OR LIMIT-TO (PUBYEAR , "2024") OR LIMIT-TO (PUBYEAR , "2025")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (OA , "all")) Date of Access: April 2026
IEEE	("autism" OR "autistic disorder" OR "autistic spectrum disorder" OR "ASD") AND ("virtual reality" OR "VR" OR "immersive technology" OR "simulated environment") AND ("user experience" OR "UX" OR "user-centered design" OR "interaction" OR "engagement") Date of Access: April 2024

B. Screening

The screening is a critical filtering step of the SLR process, which aims at narrowing down the records that had been identified during the initial stage by using preset inclusion and exclusion criteria. In this study, the rigorous screening process was implemented to determine the relevance, quality, and disciplinary suitability of the literature selected. Table II shows the selection criterion of the articles reviewed.

Besides procedural filtering, the screening judgments have substantial methodological implications and reasons. English-language publication restriction enhances interpretability and consistency in data extraction but may result in some degree of language bias by excluding potentially valuable local studies. Similarly, ensuring that the time frame remains the publications after 2020 is a way to ensure that the review incorporates the latest technological changes and current UX practices in virtual reality, which is one of the rapidly evolving spheres. Such exclusion of conference papers and reviews, possibly having omitted early-stage innovations and synthesized knowledge, supports the bias of the review toward fully developed and high-impact journal contributions. In addition, the subject areas being narrow ensure conceptual consistency and no watering down of findings by irrelevant fields. The second step of removing 12 duplicate records indicates the importance of cross-database integration and data cleaning to avoid overrepresentation of the same studies. Collectively, these screening procedures lead to enhanced internal validity, transparency and reproducibility of the review, not to mention that they are an indicator of purposeful strike of inclusiveness and methodological rigor, which are all necessary in producing high-quality results of an SLR.

TABLE II. THE SELECTION CRITERION IS SEARCHING

Criterion	Inclusion	Exclusion
Language	English	Non-English
Timeline	2020-2025	<2020
Literature Type	Journal (Article)	Conference, Book, Review
Publication Stage	Final	In Press

C. Eligibility

The eligibility stage further narrows down the filtered records by evaluating their substantive conformity to the study aims and obtaining full data to analyze. Among the 264 articles that were evaluated in terms of full-text analysis, only a significant portion of them (n=235) were eliminated according to the clearly defined criteria, such as articles that were not relevant to the research area in general, abstracts that were not related to the main objectives of the study, and those articles that were not accessible in terms of full-text assessment. This step acts as a rather important quality control gateway whereby only the studies that are directly conceptually and empirically related to the user experience in virtual reality in the context of autism are left. As a result, the resulting corpus was reduced to 29 studies that could be included in qualitative synthesis, which is a very selective but focused corpus.

D. Quality Assessment

To evaluate the methodological rigor of included studies and distinguish between high-quality and low-quality evidence, a quality assessment was conducted using the Mixed Methods Appraisal Tool (MMAT) version 2018. The MMAT is a validated instrument designed to appraise the methodological quality of five categories of studies: qualitative research, randomized controlled trials, non-randomized studies, quantitative descriptive studies, and mixed methods studies.

Each study was independently assessed by two reviewers using the appropriate MMAT criteria for its study design. The MMAT uses a set of screening questions followed by category-specific quality criteria. Studies are not assigned an overall numerical score; instead, the detailed responses to each criterion provide a comprehensive quality profile.

1) Quality assessment results of the 29 included studies:

- 8 studies (27.6%) met all applicable MMAT quality criteria, indicating high methodological rigor.
- 15 studies (51.7%) met most criteria with minor limitations.
- 6 studies (20.7%) had moderate methodological limitations.

No studies were excluded based on quality assessment; however, the quality ratings informed the interpretation and synthesis of findings. Studies with higher methodological rigor were given greater weight in the critical analysis, and limitations of lower-quality studies were explicitly noted in the thematic synthesis. The quality assessment results are summarized in Table III.

TABLE III. QUALITY ASSESSMENT SUMMARY USING MMAT

Quality Category	Number of Studies	Percentage	Key Characteristics
High Quality	8	27.6%	Met all applicable MMAT criteria; rigorous design, clear methodology, appropriate analysis
Moderate-High Quality	15	51.7%	Met most criteria with minor limitations; generally sound methodology with small gaps
Moderate Quality	6	20.7%	Notable methodological limitations; findings interpreted with caution

Common methodological strengths across high-quality studies included clear research questions, appropriate study designs, detailed descriptions of VR systems and interventions, validated outcome measures, and transparent reporting of limitations. Common weaknesses in moderate-quality studies included small sample sizes without justification, lack of control or comparison groups, limited generalizability, and insufficient detail on data analysis.

E. Data Abstraction and Analysis

The approach used in this research is an integrative analysis as it is one of the assessment methods that a researcher employs to conduct a systematic examination and synthesis of the results of varied qualitative research designs. The main aim of this strategy was to determine the overall themes and sub-themes that apply to Enhancing User Experience in Virtual Reality among Autistic children. The process started with the collection of data that formed the basis for the development of themes. The authors performed a thorough and repeated analysis of 29 chosen publications, deriving meaningful statements and evidence that were directly relevant to the study focus, as shown in Fig. 1. The studies were critically reviewed regarding their methodology and main findings to get a full picture of the existing research in the field.

The authors then held joint discussions to elaborate and develop themes based on the evidence retrieved. During the analysis, a systematic log was kept to record interpretations, emerging ideas, uncertainties, and analytical thoughts which would promote transparency and traceability. The authors constantly cross-tabulated the results of the studies in order to detect and eliminate inconsistencies in developing themes. Whenever there was a disagreement, consultations were held among the authors until they reached an agreement. The last themes were further developed to make them coherent and consistent. In order to enhance the validity and domain relevance of the thematic analysis, the identified themes were reviewed by two experts by evaluating the clarity, conceptual relevance, and consistency of the thematic categorization. Their feedback contributed to refining the analytical structure and improving the rigor and reliability of the thematic synthesis.

This study is guided by the following research questions.

1) How does virtual reality-based intervention influence user experience and skill development among children with autism in immersive learning environments?

2) How do multimodal data and evaluation techniques in virtual reality environments support the assessment of user experience and behavioral responses in children with autism?

3) What design principles and framework characteristics contribute to enhancing user experience in inclusive virtual reality systems for children with autism?

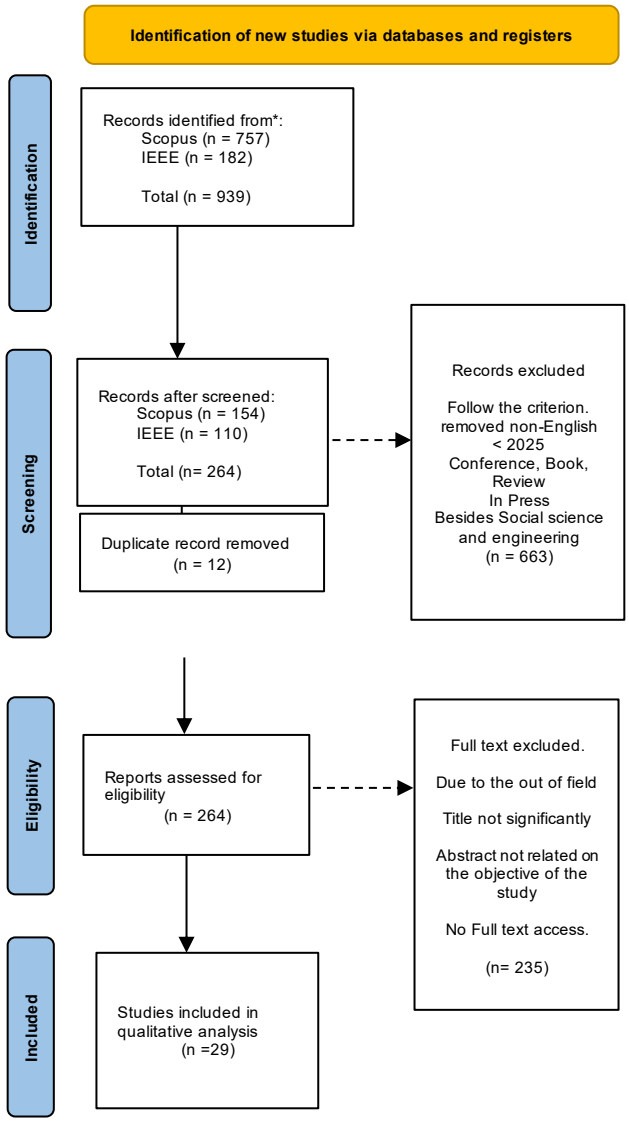


Fig. 1. PRISMA flow diagram of the proposed search study.

III. RESULTS AND FINDINGS

The following section presents the results of the systematic literature review according to the 29 selected studies that met the inclusion and quality assessment criteria. The results are divided into two broad sections: 1) an overview of the publication trends and the extracted properties of the reviewed articles, and 2) thematic analysis of the theme Enhancing user experience in Virtual Reality for Children with Autism.

A. Publication Trend and Synthesized Overview

1) Publication trend by year: Fig. 2, illustrating the temporal distribution of the reviewed articles, depicts that there is an overall increasing tendency of research output between 2020 and 2025 with several fluctuations. Publications began to decrease in 2020 (n=5), 2021 (n=3), and increased in 2022 (n=6) and 2023 (n=5), with a slight decrease in each year. It has a stronger upward trend with a greater increase in 2024 (n=9) and a high level of increase in 2025 (n=12). This trend is indicative of an increase in academic interest in integrating virtual reality (VR), user experience and interventions based on autism. The initial variability of 2020-2023 could be explained by the exploratory character of initial research and external factors like the COVID-19 pandemic that disrupted research and access to experimental conditions. This sudden increase as of 2024 will probably be fuelled by the fast development in VR technology such as more affordable and accessible hardware, and the growing recognition of the significance of personalized and immersive interventions in children with autism. In addition, the increased funding of research and interdisciplinary work in fields like digital health and assistive technologies could have led to this increase. On the whole, the upward trend indicates that the sphere transforms into less of an emerging field and more of a more mature and established one, where the focus is more on evidence-based design, user-centered solutions, and multimodal interaction in VR systems to support children with autism.

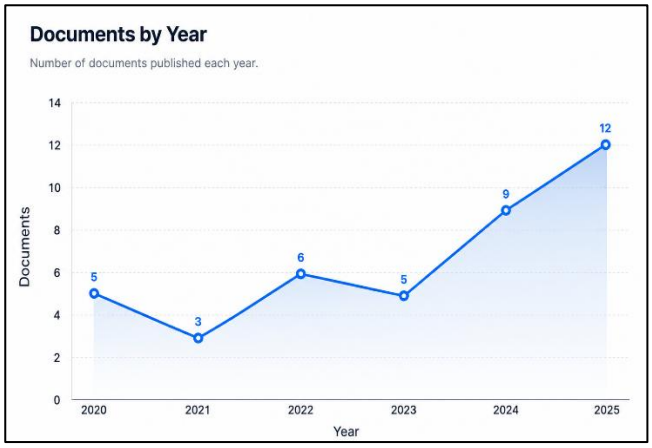


Fig. 2. Documents by year.

2) Documents by author: As Fig. 3 shows, the articles finally selected were distributed relatively widely and diffusely in terms of authorship, and the contribution of several authors was repeated quite a few times. It is interesting to note that Jung, S., Mei, C., and Moon, J. were the most frequent authors with two publications each, and the other authors had a single publication. This means that no single or small group of scholars dominates the field, but it is characterized by a large range of contributors. This may be because of the presence of recurrent authors, as this may be interpreted as a sign of nascent collaborative organizations or research partnerships, potentially being interdisciplinary collaboration between areas such as virtual reality, human-computer interactions, and autism research. The reason why these partnerships are so important is that they merge technical innovation and clinical and consumer-based perspectives. Overall, the trend is that the research area is still forming and that the number of contributors is growing but has not yet solidified, which implies that it is an emerging field where various contributions are being made to enhance research development, but it may be required to have more solid networks of collaboration to enhance research maturity and sustainability.

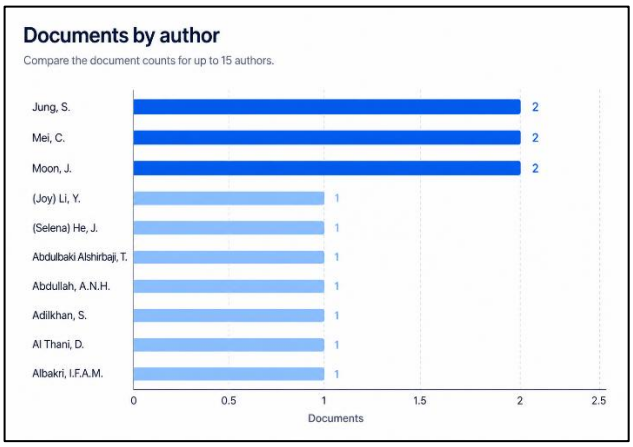


Fig. 3. Documents by author.

3) Documents by country: Fig. 4 demonstrates that the geographical pattern of the identified studies indicates a high level of concentration of the research production in several most productive countries, the United States (n=17) is the most productive country, and China (n=4) comes next. Other nations have a moderate contribution (n=3 each on average) with Argentina, Greece, Italy, and Norway being underrepresented (n=2 each). The supremacy of the United States is explained by the abundance of technological infrastructure, the high level of research investments, and the developed healthcare and educational systems in which the emphasis is put on the innovative interventions towards the developmental disorders. Likewise, the rise in contribution of China is indicative of its rise in investment in new technology like VR and digital health. By comparison, the smaller and more disjointed input of the Malaysian and other Asian nations indicates that, although interest in the region is increasing, research activities are less organized and could be limited by a lack of funds, infrastructural differences, and differences in healthcare priorities. On the whole, the field proves to be geographically uneven although it is geographically participatory with a prevalence in developed areas. This disparity explains why increasing international cooperation and capacity building is essential to make more inclusive and context-dependent improvements, especially in those areas that are underrepresented, and VR-based interventions may be of great assistance to children with autism.

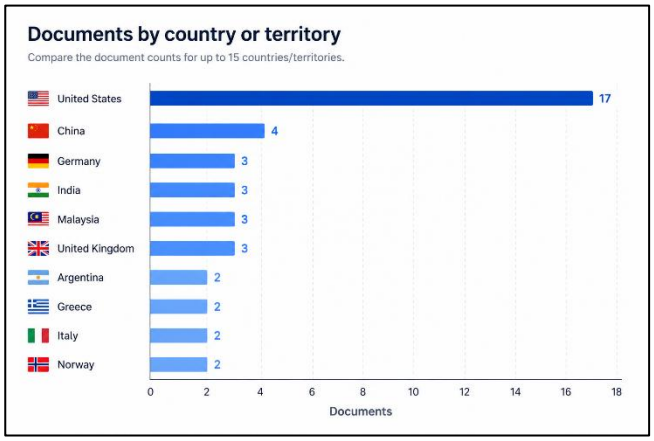


Fig. 4. Documents by country or territory.

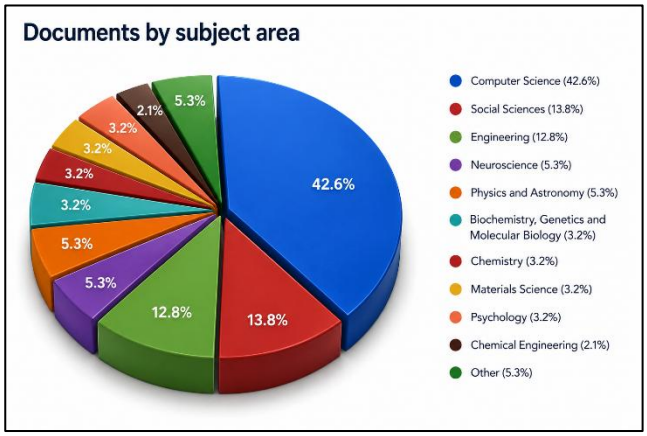


Fig. 5. Documents by subject area.

4) Documents by subject area: As Fig. 5 demonstrates, the choice of the articles by subject area reveals that computer science (42.6 percent) is highly prevalent, indicating that the creation of virtual reality (VR) technologies and the design of interactions are the most fundamental basis of this research area. This is followed by social sciences (13.8%), engineering (12.8%), which specializes in user-centered design, usability, and system development. Other smaller groups are neuroscience (5.3%), physics and astronomy (5.3%), and shares of psychology, biochemistry, chemistry, materials science and chemical engineering (all about 23). The value of cognitive, behavioral and sensory processing attributes in children with autism is indicative of psychology, neuroscience, engineering and physical sciences, which contribute to the hardware engineering, simulation and optimization of the systems. This interdisciplinary spread shows that the enhancement of user experience in VR in children with autism cannot be attributed to a single domain but instead, it requires the use of technological innovation in combination with clinical, behavioral and social strategies. Such collaboration enables the development of more detailed and effective solutions, as computer scientists are involved in the development of immersive systems, psychologists and the sphere related to healthcare ensure the relevance and usefulness of the interventions, and social sciences advise us on usability and accessibility issues. Overall, the overlap of the different areas contributed to the improvement of the research as it enables the creation of holistic design solutions, makes VR interventions more relevant and applicable, and allows taking a closer look at the ways of immersive technologies usage in addressing the needs of children with autism.

B. Thematic Findings

This section reviews the findings of the systematic literature review on the research questions and places them in the wider context of Enhancing User Experience in Virtual Reality with Autistic Children. These results highlight the main contributions, limitations, and future research and practice implications. Table IV presents the selected studies categorized according to the identified themes.

TABLE IV. THEMATIC CATEGORIZATION OF SELECTED STUDIES

Themes	#	Article Title	Ref
Theme 1: VR-Based Intervention and Skill Development for Autism	1	An Exploratory Analysis of Increasing Self-Efficacy of Adults with Autism Spectrum Disorder through the Use of Multimedia Training Stimuli	[35]
	2	Impact of virtual reality learning environments on skills development in students with ASD	[36]
	3	Combining Virtual Reality Visual Novels and Social Stories to Support Social and Emotional Development in Children with Autism Spectrum Disorder	[37]
	4	Charisma™ virtual social training: A digital health platform and protocol	[38]
	5	Development of a Social Communication Skills Intervention Using Video Modeling and Spherical Video-Based Virtual Reality for High Functioning Autism Spectrum Disorder Youth: A Preliminary Study	[39]
	6	"Can I be More Social with a Chatbot?": Social Connectedness Through Interactions of Autistic Adults with a Conversational Virtual Human	[40]
	7	Enhancing attention in autism spectrum disorder: comparative analysis of virtual reality-based training programs using physiological data	[41]
	8	Using virtual reality to enhance attention for autistic spectrum disorder with eye tracking	[42]
	9	Designing Gamified Virtual Reality Intervention Based on Experiential Learning to Enhance Social Reciprocity in Children with Autism Spectrum Disorder	[43]
	10	Learning experience design of verbal prompts in virtual reality-based training for autistic children	[44]
	11	Serious Game for VR Road Crossing in Special Needs Education	[45]

	12	Virtual Reality Air Travel Training with Children on the Autism Spectrum: A Preliminary Report	[46]
	13	FECTS: A Facial Emotion Cognition and Training System for Chinese Children with Autism Spectrum Disorder	[47]
	14	SoundFields: A virtual reality game designed to address auditory hypersensitivity in individuals with autism spectrum disorder	[48]
Theme 2: Assessment, Evaluation, and Multimodal Analytics in VR for Autism	15	Semi-Supervised Behavior Labeling Using Multimodal Data during Virtual Teamwork-Based Collaborative Activities	[49]
	16	Virtual Reality-Based Approach to Evaluate Emotional Everyday Scenarios for a Digital Health Application	[50]
	17	Applying multimodal data fusion to track autistic adolescents' representational flexibility development during virtual reality-based training	[51]
	18	On the Evaluation of Engagement in Immersive Applications When Users Are on the Autism Spectrum	[52]
	19	A Virtual Reality-Based System for the Screening and Classification of Autism	[53]
	20	ILAT (Software as a Service): Interactive Learning Application Tool for Autism Screening and Assessment in children with Autism Spectrum Disorder	[54]
	21	A Multimodal Framework for Integrating Motion Capture Monitoring into ASD Screening Utilizing XR Applications	[55]
	22	Individual Variability in Cognitive Engagement and Performance Adaptation During Virtual Reality Interaction: A Comparative EEG Study of Autistic and Neurotypical Individuals	[56]
Theme 3: Design, Framework, and Inclusive VR Systems for Autism	23	AutiSim: A Virtual Reality Simulation Game Based on the Autism Spectrum Disorder	[57]
	24	Law enforcement training using simulation for locally customized encounters	[58]
	25	"Someone who has ADHD or someone who has autism should make the Rules": A participatory study of neurodivergent Child perspectives on the ethics of Extended reality technologies	[59]
	26	Understanding experiences and interactions of children with Asperger's syndrome in Virtual Reality-based learning systems	[60]
	27	Inclusive Learning through Immersive Virtual Reality and Semantic Embodied Conversational Agent: A case study in children with autism	[61]
	28	Daily Life Adaptation in Autism: A Co-Design Framework for the Validation of Virtual Reality Experiential Training Systems	[62]
	29	Virtual Reality as a Tool for Children with Autism Spectrum Disorder	[63]

1) Theme 1: VR-based intervention and skill development for autism.

The results all point to a positive impact of VR-based interventions on cognitive, behavioural, and self-efficacy measures in people with autism, especially in cases of immersive and multimedia applications. The results in [35] indicate that there is an objectively observed change in self-efficacy and an increase in engagement during vocational training, indicating that the interactive settings can be used to boost confidence in performing tasks. Equally, [36], [43] note enhancement in cognitive, social communication and behavioural adaptation in structured VR learning situations, yet statistical significance is sensitive to sample size and design limitations. Additional findings [45], [46] indicate that VR simulation allows effective transfer of acquired skills to real-life situations, especially in everyday living activities like travel and road safety, which adds to the ecological validity of VR interventions.

There is also a high level of convergence in the outcomes of analyzing social communication and emotional development, whereby VR-based tools offer controlled but realistic practice environments. [38] note that structured virtual social training results in significant emotion recognition, social inferencing, and self-perception improvements, whereas [39] highlight that video modelling and spherical VR are particularly effective at understanding due to multi-perspective learning. Multiple studies reported that immersive and interactive VR systems improved emotional recognition and expression among autistic users [37], [47]. Moreover, [40] show that conversational virtual agents can facilitate perceived social connectedness, but anticipations of stronger interaction are not yet achieved, meaning that emotional realism and responsiveness continue to be developing issues in VR-based social systems.

Another prevalent finding of the studies reviewed is attention enhancement, especially when VR is used in combination with physiological monitoring and adaptive feedback processes. Several studies reported that reinforcement strategies and physiological monitoring enhanced attentional engagement within VR environments [41, 64]. [65] also indicates that thoughtfully created verbal cues in VR have the potential to affect the patterns of interaction and maintain the attention of users when performing training activities. These results are consistent with [48] in which exposure-based VR gaming minimized anxiety connected to auditory hypersensitivity, which indirectly underpins attentional stability. Taken together, these findings indicate that individual and modular VR processes are imperative in the treatment of attentional variability in autism.

Although the results are encouraging, some limitations and design aspects are consistently pointed out in research, which means that VR interventions should be refined further. Most research papers, such as [66], [67] have limitations of small sample size, limited intervention period, and limitations of generalization of results. Improved interactivity, user adaptability, and the integration of external support systems,

including parental involvement, are also highlighted in [39] and [65]. Moreover, [40] emphasizes the discrepancy between the expectations of the user and the technological possibilities of conversational systems. The overall findings of these observations demonstrate that, although VR-based interventions have great potential in improving the user experience and skill development of children with autism, future studies should focus on addressing scalability, long-term assessment, and user-related design aspects.

While the majority of studies demonstrated positive outcomes from VR-based interventions, the magnitude of effectiveness varied depending on immersion level, interaction modality, and participant characteristics. Studies utilizing adaptive and gamified VR systems generally demonstrated stronger improvements in social reciprocity and attentional engagement compared to static or less interactive environments. However, contradictions also emerged across the literature. For example, while immersive HMD-based systems increased engagement and realism, several studies simultaneously reported concerns regarding sensory overload and cybersickness among highly sensitive autistic users. Similarly, conversational virtual agents improved perceived social connectedness, yet emotional realism and natural responsiveness remained limited in more complex interactions. These findings suggest that VR effectiveness is not solely dependent on technological sophistication but rather on the extent to which the environment aligns with individual sensory and cognitive profiles.

2) Theme 2: Evaluation and multimodal analytics in VR for autism.

Multimodal data integration is an important technique to improve the accuracy and efficiency of behaviour recognition and evaluation in VR-based autism systems. Assessment, Evaluation, and Multimodal Analytics in VR on Autism report that semi-supervised learning can be a useful method to reduce the amount of manual annotation work but still achieve high prediction accuracy, which suggests that an efficient method to monitor behaviour in real-time is practical. Likewise, [68] demonstrate that using a combination of behavioural, physiological, and interaction data enhances predictive accuracy in monitoring cognitive skill development, especially representational flexibility. [69] also support the idea that the combination of subjective assessment with psychophysiological measures like ECG and EDA will enable the distinction between the complexity of a scenario in VR environments. All these results indicate that multimodal analytics is a more reliable and holistic representation of user states than single-source data solutions.

The other significant result pertains to the creation of objective and automated measures of autism screening and classification via VR technologies. [70] show that behavior monitoring in simulated social interactions is amenable to machine learning methods to differentiate between autistic and normal individuals with a high degree of accuracy, which proves the practicability of VR-based diagnostic devices. Also [54] emphasize the importance of integrated digital platforms, which are a combination of VR, artificial intelligence, and cloud-based systems that can aid early screening, behavioural monitoring, and intervention planning. Moreover, [55] note that to enhance the quality of behavioral information gathered during the assessment, it is crucial to have very accurate motion capture and full-body tracking in XR settings. Such systems are indications of a transition to more objective, data-driven approaches to lessening dependence on subjective clinical judgment.

The evaluation of the degree of involvement and mental responses in VR settings also emerges as a major point of interest in the findings. [71] demonstrate that immersive VR applications can potentially generate high engagement levels among autistic users, and the conventional evaluation measures may have to be redefined due to the unique behavioural aspects. Moreover, [56] reveal that cognitive engagement measured by EEG signals differs significantly based on certain individual factors such as IQ, familiarity with technologies, and pre-exposure to VR. [50] also, demonstrate the objectivity of emotional and physiological reactions in order to test the engagement of the user with varying levels of complexity in scenarios. These experiments lead to the fact that not all VR engagement is identical and that there ought to be adaptive evaluation models, which consider individual differences.

Regardless of these developments, various difficulties and constraints are always reported, especially with regard to standardisation, scalability, and personalisation of assessment models. According to [72] and [68], machine learning models demonstrate high performance, but their effectiveness is limited by the data quality and diversity, which are still limited on a small scale. [56] also emphasise that the variability between individuals makes it harder to generalise the results to other populations, and it requires more customised system design. Moreover, [71] emphasize the absence of agreement on assessment models that should be used with autistic users whereas [55] note the technical issues surrounding the necessary motion tracking.

The reviewed studies collectively demonstrate a transition from subjective observational assessment toward data-driven and multimodal evaluation frameworks. Nevertheless, important methodological differences remain evident across studies. While machine learning models achieved promising classification accuracy, performance was heavily dependent on data quality, sample diversity, and sensor reliability. Studies employing EEG, eye-tracking, and physiological monitoring reported stronger predictive capabilities, yet these approaches also introduced challenges related to equipment accessibility, calibration complexity, and scalability in real-world settings. Consequently, although multimodal analytics improve behavioral interpretation and personalization, the lack of standardized evaluation frameworks continues to limit comparability across VR autism studies.

3) Theme 3: Design, framework, and inclusive VR systems for autism.

The results indicate that the design of VR systems is becoming more oriented towards simulation-based practices to enhance awareness and insights into the experiences of autism. [56] show that immersive simulation games are also capable of capturing sensory sensitivities and social challenges, and usability tests show that the level of user comfort relating to simulator sickness is generally acceptable. On the same note,

[58] demonstrate that VR-based training environments offer realistic, but controlled practice environments, especially in the case of professional training such as law enforcement, where exposure to complex situations must be repeated. [73] also disclose that children with AS have been reported to develop an ability to adapt to virtual characters with time as they engage with VR learning systems, implying that well-thought-out virtual environments can elicit behavioral engagement in the long run. All these results suggest that the important design factors in VR systems that should be considered in autism-related uses include realism, immersion, and contextual relevance.

The other significant point that the findings raise is the focus on inclusive and assistive design to improve user experience and accessibility. According to [74], the combination of conversational agents and immersive VR systems leads to educational involvement as well as social integration, especially via natural language interaction. It is also pointed out in [75] that VR with virtual assistants has the potential to increase attention span and offer personalised assistance in everyday learning scenarios. Moreover, [76] reveal that co-design models with caregivers lead to system usability and user satisfaction and it is crucial to note that collaborative development processes are highly valued. All these studies suggest that inclusive VR systems must have adaptive communication and stakeholder engagement to ensure that children with autism engage in a significant manner.

Ethical considerations and participatory design is another valuable insight in the development of VR systems that can be used by neurodivergent users. [59] emphasize that children with autism insist that they are consulted into the design process and are concerned about the privacy of their data, overexposure, and potential misuse of immersive technologies. [58] indirectly support this line of thought and conclude that it is necessary to modify VR situations to its own user conditions to ensure relevance and acceptability. [57] also note the challenge of covering the experiences of autism in a manner that is not perpetuating stereotypes, which means that one has to be ethically mindful when developing content. Such findings reveal that technological innovation and the interests of users must be reconciled by adopting user-centred and ethically sound design practices.

In spite of such improvements, some limitations and design problems are always documented in research. According to [73] and [76], early pain and the necessity of a mentor are still obstacles to certain users, especially in their early exposure to VR systems. It is also noted by [75] that the variability in attention and engagement demands adaptive systems that can respond to the needs of individuals. In addition, [74] emphasize that even though promising results are demonstrated by immersive systems, the scalability and the effectiveness in the long run are under-researched. These results indicate that the design of VR systems in the future must focus more on flexibility, long-term usability analysis, and feedback mechanisms.

Across the reviewed literature, inclusive and participatory design consistently emerged as a critical factor influencing usability and long-term engagement. Studies involving caregivers, therapists, educators, and autistic users during development generally reported higher acceptance and improved interaction quality compared to systems developed without stakeholder collaboration. However, tensions remain between immersive realism and sensory comfort. While realistic VR simulations improved contextual learning and emotional engagement, excessive sensory stimulation occasionally reduced usability for participants with hypersensitivity. These findings indicate that future VR systems should prioritize adaptive sensory regulation and personalized interaction settings rather than relying solely on increased immersion.

IV. Discussion and Conclusion

The proposed systematic literature review is designed to explore the effectiveness of virtual reality (VR) technologies in enhancing the user experience among children with autism spectrum disorder through a synthesis of recent research published in 2020-2025.

The results indicate that there are a number of patterns that are common to the chosen studies. First, VR-based interventions have a high potential in enhancing social communication, cognitive skills, emotional recognition, and daily living skills in children with autism. These enhancements are more noticeable when immersive and interactive elements are integrated into structured learning environments. Second, the multimodal data system, including that based on behavioural, physiological, and interaction aspects, proves to be a valuable contribution to the field of user experience measurement. These approaches enable objective, real-time, and personalised assessment compared to conventional approaches. Third, the principles of user-centred and inclusive design are defined as the most important ones due to the enhanced engagement and usability. The adaptive sensory control, intuitive interaction techniques, co-design participation and ethical issues are some of the factors that are important in the development of positive user experiences. Artificial intelligence, along with real-time feedback systems, are also gaining popularity in the literature to enable personalisation and dynamic adaptation of VR systems.

The present review contributes to the area because it provides a methodical synthesis of a disjointed literature and brings all the various views together in three coherent themes: VR-based intervention and skill development, multimodal assessment and analytics, and inclusive design frameworks. This classification gives the review a more conceptual understanding of how the technological, behavioural and design aspects intersect to enhance user experience among children with autism. Furthermore, the review highlights a gradual shift from technology-centered approaches toward more human-centered, adaptive, and ethically conscious VR systems. The review also contributes to existing knowledge by emphasizing the importance of addressing sensory, emotional, and cognitive aspects of VR design rather than focusing solely on functional outcomes.

These findings have significant practical implications for different stakeholders. VR systems may serve as supportive tools in educational and clinical settings for skill training, behavioural therapy, and experiential learning, particularly for individuals with diverse sensory profiles. It is possible to add multimodal analytics that may assist practitioners in monitoring

progress in a more objective manner and developing personalised interventions. The findings are a reminder to the developers and designers that inclusive and user-centred design should be adopted, and VR applications must be made accessible, adaptable, and responsive to the diverse needs of autistic users.

Though these contributions are significant, several limitations remain evident across the reviewed studies. Most studies relied on relatively small participant groups, short intervention durations, and limited demographic diversity, reducing the generalizability of findings across broader autism populations. Future longitudinal research should therefore investigate the long-term effectiveness of VR interventions over periods ranging from six months to several years to better understand behavioral retention, emotional adaptation, and sustained engagement. For example, longitudinal studies conducted within schools or therapy centers could evaluate whether repeated VR-based social communication training contributes to measurable improvements in classroom interaction, emotional regulation, or independent living skills over time.

Future studies should also include more diverse participant groups across age ranges, cultural backgrounds, socioeconomic conditions, and autism support levels. Cross-cultural investigations are particularly important because sensory preferences, communication styles, educational practices, and caregiver expectations may differ significantly between countries and regions. For instance, VR learning environments developed for Western educational settings may require adaptation before implementation in Asian, Middle Eastern, or rural educational contexts where learning structures and technology accessibility differ substantially.

Another important direction involves the development of standardized evaluation frameworks for VR-based autism interventions. Current studies frequently employ different usability metrics, engagement scales, and physiological indicators, making direct comparison difficult. Future frameworks should integrate behavioral observation, eye-tracking, physiological monitoring, cognitive performance measures, and user satisfaction into unified assessment models. Such standardization would improve reproducibility and allow researchers to compare intervention outcomes across institutions and populations more effectively.

The integration of advanced technologies such as artificial intelligence, adaptive systems, and machine learning also presents promising opportunities for real-world implementation. AI-driven VR systems could dynamically adjust environmental complexity, sound intensity, visual stimulation, and task difficulty based on real-time behavioral and physiological responses from autistic users. In healthcare settings, clinicians may utilize adaptive VR systems for exposure therapy, emotional recognition training, or anxiety management. In educational environments, teachers could employ personalized VR simulations to support social interaction, classroom transition preparation, and vocational skill development. Similarly, organizations involved in workforce readiness or public services may adopt VR simulations to prepare autistic individuals for transportation use, workplace communication, or emergency response situations. Despite these opportunities, future implementation must also address ethical concerns related to privacy, data security, informed consent, and overdependence on immersive technologies. Participatory and user-centered design approaches should therefore remain central to future VR development to ensure that autistic individuals, caregivers, therapists, and educators actively contribute to system design and decision-making processes.

To conclude, this systematic review demonstrates the increased significance of virtual reality as a means of improving user experience in children with autism, but also reveals the important gaps that need to be addressed through future research. The integration of existing knowledge shows that the best VR systems should be based on technological innovations and user-centred design, multimodal assessment, and ethical considerations. Systematic reviews in this field play a critical role in systematising the rapidly growing body of knowledge, reducing fragmentation, and guiding future research directions. This type of evidence-based synthesis can not only advance theoretical knowledge but also support the development of more effective, inclusive, and sustainable VR solutions for individuals with autism.

ACKNOWLEDGMENT

This research was supported by the Ministry of Higher Education (MoHE) Malaysia under the Research Consortium Grant 2023 (Reference No: RK059). The authors gratefully acknowledge Universiti Sultan Zainal Abidin (UniSZA) for institutional support and project coordination. Appreciation is also extended to all consortium members and research collaborators for their valuable contributions to this study.

REFERENCES

[1] N. S. Mohd Shapri and N. L. Rashid, "Exploring Prehistoric Era in Virtual Reality Walkthrough," International Journal of Engineering Trends and Technology, vol. 72, no. 6, pp. 1–7, Jun. 2024, doi: 10.14445/22315381/IJETT-V72I6P101.

[2] M. Danish et al., "Validate the Users' Comfortable Level in the Virtual Reality Walkthrough Environment for Minimizing Motion Sickness." [Online]. Available: www.ijacsa.thesai.org

[3] A. Koirala, Z. Yu, H. Schiltz, A. Van Hecke, B. Armstrong, and Z. Zheng, "A Preliminary Exploration of Virtual Reality-Based Visual and Touch Sensory Processing Assessment for Adolescents with Autism Spectrum Disorder," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 619–628, 2021, doi: 10.1109/TNSRE.2021.3064148.

[4] M. N. M. Nushkan, W. D. Sumadhya, A. R. Shaveesha, K. Manathunga, D. Chamara, and I. Gamage, "Effectiveness of Virtual Realityin Improving Learning Outcomes for Children with Autism," in 2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/I2CT61223.2024.10543652.

[5] J. Zhao et al., "Virtual reality technology enhances the cognitive and social communication of children with autism spectrum disorder," Front. Public Health, vol. 10, 2022, doi: 10.3389/fpubh.2022.1029392.

[6] N. A. Bakar, M. S. Ramli, S. Sulaiman, and Y. Hussain, "Perceived usefulness of augmented reality application as educational intervention for children with autism spectrum disorder," Multidisciplinary Science Journal, vol. 8, no. 2, 2026, doi: 10.31893/multiscience.2026085.

[7] N. A. Bakar, M. S. Ramli, S. Sulaiman, N. Zamri, W. P. I. Wan, and R. Muhamad, "Impact assessment of perceived efficacy and training on perceived usefulness of augmented reality application as educational intervention among ASD children," Multidisciplinary Science Journal, vol. 8, no. 2, 2026, doi: 10.31893/multiscience.2026083.

[8] V. Bhagat, M. Haque, N. Bin Simbak, and R. Husain, "Emotional Regulation in Autism Spectrum Disorders: A New Proposed Model for Regulating Emotions through Parent Education," 2009. [Online]. Available: www.jgpt.co.in

[9] A. E. Keys, O. Oyekoya, and D. Chan, "Exploring Effects of Interactive Virtual Reality Sensory Environment on Anxiety Reduction in Adolescents with Autism," in ISS Companion 2024 - Companion Proceedings of the Conference on Interactive Surfaces and Spaces, B. Ens and S. Carpendale, Eds., Association for Computing Machinery, Inc, 2024, pp. 48–53. doi: 10.1145/3696762.3698051.

[10] Z. Yu, S. Iadarola, S. Daley, and Z. Zheng, "A Screen-Based Multimodal Virtual Classroom Interface for Understanding Behavioral Sensory Responses in Autistic Adolescents: A Pilot Study," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 33, pp. 4067–4077, 2025, doi: 10.1109/TNSRE.2025.3619484.

[11] M. L. Max and J. C. Burke, "Virtual reality for autism communication and education, with lessons for medical training simulators," in Studies in Health Technology and Informatics, IOS Press, 1997, pp. 46–53. doi: 10.3233/978-1-60750-883-0-46.

[12] B. Güçlü, C. Tanidir, N. M. Mukaddes, and F. Ünal, "Tactile sensitivity of normal and autistic children," Somatosens. Mot. Res., vol. 24, no. 1–2, pp. 21–33, 2007, doi: 10.1080/08990220601179418.

[13] M. Diepman and N. Brady, "Sensory responsivity and its relation to alexithymia, social processing and restricted interests and repetitive behaviour in autistic children," Res. Autism Spectr. Disord., vol. 118, 2024, doi: 10.1016/j.rasd.2024.102470.

[14] L. Maddalon, M. E. Minissi, S. Cervera-Torres, A. Hervás, S. Gómez-García, and M. Alcañiz, "Multimodal Interaction in ASD Children: A Usability Study of a Portable Hybrid VR System," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), M. Antona and C. Stephanidis, Eds., Springer Science and Business Media Deutschland GmbH, 2023, pp. 614–624. doi: 10.1007/978-3-031-35681-0_40.

[15] J. F. Herrero and G. Lorenzo, "An immersive virtual reality educational intervention on people with autism spectrum disorders (ASD) for the development of communication skills and problem solving," Educ. Inf. Technol. (Dordr)., vol. 25, no. 3, pp. 1689–1722, 2020, doi: 10.1007/s10639-019-10050-0.

[16] E. Bekele et al., "Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD)," in Proceedings - IEEE Virtual Reality, V. Interrante, T. Hollerer, E. Suma, and A. Lecuyer, Eds., IEEE Computer Society, 2016, pp. 121–130. doi: 10.1109/VR.2016.7504695.

[17] Y. Zhao et al., "P-MARS: Design of a VR-Based Ergotherapy System for Children with Autism and its Longitudinal Tracking Evaluation," in Proceedings - 2025 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2025, U. Eck, G. Lee, A. Plopski, M. Smith, Q. Sun, and M. Tatzgern, Eds., Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1534–1542. doi: 10.1109/ISMAR67309.2025.00157.

[18] M. Schmidt et al., "Toward a strengths-based model for designing virtual reality learning experiences for autistic users," Autism, vol. 28, no. 7, pp. 1809–1827, 2024, doi: 10.1177/13623613231208579.

[19] L. Boyd, B. Wasserman, G. Hayes, K. Day, K. Abdo, and E. Linstead, "Paper prototyping comfortable VR play for diverse sensory needs," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, 2019. doi: 10.1145/3290607.3313080.

[20] V. Bauer, T. Bouchara, O. Duris, C. Labossière, M.-N. Clément, and P. Bourdot, "Evaluating the Acceptability and Usability of a Head-Mounted Augmented Reality Approach for Autistic Children with High Support Needs," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), G. Zachmann, A. R. M, P. Bourdot, M. Marchal, J. Stefanucci, and X. Yang, Eds., Springer Science and Business Media Deutschland GmbH, 2022, pp. 53–72. doi: 10.1007/978-3-031-16234-3_4.

[21] N. Newbutt and R. Bradley, "Using immersive virtual reality with autistic pupils: moving towards greater inclusion and co-participation through ethical practices," J. Enabling Technol., vol. 16, no. 2, pp. 124–140, 2022, doi: 10.1108/JET-01-2022-0010.

[22] N. Newbutt, C. Sung, H.-J. Kuo, M. J. Leahy, C.-C. Lin, and B. Tong, "Brief Report: A Pilot Study of the Use of a Virtual Reality Headset in Autism Populations," J. Autism Dev. Disord., vol. 46, no. 9, pp. 3166–3176, 2016, doi: 10.1007/s10803-016-2830-5.

[23] A. Grabowski, "Acceptance and Usability of Different Virtual Reality-Based Technologies Supporting Physical Rehabilitation of the Upper Limb: Results of Experiments with Patients After Upper Limb Injury at Work," Games Health J., vol. 14, no. 5, pp. 389–397, 2025, doi: 10.1089/g4h.2024.0071.

[24] K. Zhang, J. Chen, and Z. Yang, "Assessment and intervention with virtual reality technology for children aged 3–12 years with autism spectrum disorders: A scoping review," Educ. Inf. Technol. (Dordr)., vol. 30, no. 14, pp. 19703–19740, 2025, doi: 10.1007/s10639-025-13561-1.

[25] M. I. Rosli, Z. Che Embi, and J. Abdullah, "A Non-immersive Virtual Reality Application for Children with Autism Spectrum Disorder," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), A. I. Cristea and C. Troussas, Eds., Springer Science and Business Media Deutschland GmbH, 2021, pp. 519–525. doi: 10.1007/978-3-030-80421-3_57.

[26] P. Mittal et al., "Effect of immersive virtual reality-based training on cognitive, social, and emotional skills in children and adolescents with autism spectrum disorder: A meta-analysis of randomized controlled trials," Res. Dev. Disabil., vol. 151, 2024, doi: 10.1016/j.ridd.2024.104771.

[27] M. A. Md Ali, M. N. Ahmad, W. S. Wan Ismail, and N. S. Mohamad Aun, "VIRTUAL REALITY APPLICATIONS IN AUTISM SPECTRUM DISORDER: A SYSTEMATIC REVIEW," 15th International Conference on ICT, Society and Human Beings, ICT 2022, 19th International Conference on Web Based Communities and Social Media, WBC 2022 and 14th International Conference on e-Health, EH 2022 - Held at the 16th Multi Conference on Comp…, 2022, Accessed: Apr. 04, 2026. [Online]. Available: https://www.scopus.com/pages/publications/85142223586?origin=scopusAI

[28] W. Fu, T. Tao, J. Wang, Z. Peng, and Y. Wang, "Application of Virtual Reality Technology in Intervention for Children With Autism Spectrum Disorder: A Systematic Review," Journal of Special Education Technology, vol. 40, no. 4, pp. 519–529, 2025, doi: 10.1177/01626434251317984.

[29] D. Cunningham and M. Krishack, "Virtual reality promotes visual and cognitive function in rehabilitation," Cyberpsychology and Behavior, vol. 2, no. 1, pp. 19–23, 1999, doi: 10.1089/cpb.1999.2.19.

[30] S.-N. Ahn, "Virtual, augmented, and mixed reality training for improving social skills in individuals with Autism Spectrum Disorder: A systematic review," Hong Kong Journal of Occupational Therapy, vol. 38, no. 2, pp. 167–178, 2025, doi: 10.1177/15691861251369043.

[31] Y. Politis, C. Sung, L. Goodman, and M. Leahy, "Conversation skills training for people with autism through virtual reality: using responsible research and innovation approach," Advances in Autism, vol. 6, no. 1, pp. 3–16, 2020, doi: 10.1108/AIA-05-2018-0017.

[32] K. Wagner et al., "From Innovation to Integration: Classroom Strategies for Using VR With Students on the Autism Spectrum," Journal of Special Education Technology, 2025, doi: 10.1177/01626434251387307.

[33] A. Carreon, A. Rowland, S. Smith, A. Lowery, and M. Mosher, "Designing for Successful Educational Experiences in Virtual Reality Environments," Journal of Special Education Technology, vol. 40, no. 3, pp. 298–313, 2025, doi: 10.1177/01626434241277191.

[34] R. M. Silva, D. Carvalho, P. Martins, and T. Rocha, "Systematic Literature Review of the Use of Virtual Reality in the Inclusion of Children with Autism Spectrum Disorders (ASD)," in Lecture Notes in Computer Science, H. Y.-M. and T. Rocha, Eds., Springer Science and Business Media Deutschland GmbH, 2023, pp. 501–509. doi: 10.1007/978-3-031-40113-8_49.

[35] G. E. Kuper, K. Ksobiech, J. Wickert, F. Leighton, and E. Frederick, "An Exploratory Analysis of Increasing Self-Efficacy of Adults with Autism Spectrum Disorder through the Use of Multimedia Training Stimuli," Cyberpsychol. Behav. Soc. Netw., vol. 23, no. 1, pp. 34–40, 2020, doi: 10.1089/cyber.2019.0111.

[36] R. M. Silva, P. Martins, and T. Rocha, "Impact of virtual reality learning environments on skills development in students with ASD," Computers and Education Open, vol. 9, 2025, doi: 10.1016/j.caeo.2025.100298.

[37] K. Atsalaki and I. Kazanidis, "Combining Virtual Reality Visual Novels and Social Stories to Support Social and Emotional Development in Children with Autism Spectrum Disorder," Applied Sciences (Switzerland), vol. 15, no. 12, 2025, doi: 10.3390/app15126584.

[38] M. Johnson, A. M. Tate, K. Tate, S. A. Laane, Z. Chang, and S. B. Chapman, "CharismaTM virtual social training: A digital health platform and protocol," Front. Virtual Real., vol. 3, 2022, doi: 10.3389/frvir.2022.1004162.

[39] M. A. F. A. Basri et al., "Development of a Social Communication Skills Intervention Using Video Modeling and Spherical Video-Based Virtual Reality for High Functioning Autism Spectrum Disorder Youth: A Preliminary Study," IEEE Access, vol. 12, pp. 77976–77987, 2024, doi: 10.1109/ACCESS.2024.3407114.

[40] A. Xygkou, P. Siriaraya, W.-J. She, A. Covaci, and C. S. Ang, "'Can I be More Social with a Chatbot?': Social Connectedness Through Interactions of Autistic Adults with a Conversational Virtual Human," Int. J. Hum. Comput. Interact., vol. 40, no. 24, pp. 8937–8954, 2024, doi: 10.1080/10447318.2023.2292880.

[41] B. S. Sanku, Y. Li, S. Jung, C. Mei, and J. He, "Enhancing attention in autism spectrum disorder: comparative analysis of virtual reality-based training programs using physiological data," Front. Comput. Sci., vol. 5, 2023, doi: 10.3389/fcomp.2023.1250652.

[42] R. Razzak, Y. (Joy) Li, J. (Selena) He, S. Jung, C. Mei, and Y. Huang, "Using virtual reality to enhance attention for autistic spectrum disorder with eye tracking," High-Confidence Computing, vol. 5, no. 1, 2025, doi: 10.1016/j.hcc.2024.100234.

[43] P. Julrode, P. Worragin, P. Ariya, K. Puritat, and K. Intawong, "Designing Gamified Virtual Reality Intervention Based on Experiential Learning to Enhance Social Reciprocity in Children with Autism Spectrum Disorder," Educ. Sci. (Basel)., vol. 15, no. 9, 2025, doi: 10.3390/educsci15091104.

[44] J. Moon, "Learning experience design of verbal prompts in virtual reality-based training for autistic children," Research in Learning Technology, vol. 32, 2024, doi: 10.25304/rlt.v32.3129.

[45] Q. P. Tan, L. Huang, D. Xu, Y. Cen, and Q. Cao, "Serious Game for VR Road Crossing in Special Needs Education," Electronics (Switzerland), vol. 11, no. 16, 2022, doi: 10.3390/electronics11162568.

[46] I. T. Miller, B. K. Wiederhold, C. S. Miller, and M. D. Wiederhold, "Virtual Reality Air Travel Training with Children on the Autism Spectrum: A Preliminary Report," Cyberpsychol. Behav. Soc. Netw., vol. 23, no. 1, pp. 10–15, 2020, doi: 10.1089/cyber.2019.0093.

[47] G. Wan et al., "FECTS: A Facial Emotion Cognition and Training System for Chinese Children with Autism Spectrum Disorder," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9213526.

[48] D. Johnston, H. Egermann, and G. Kearney, "SoundFields: A virtual reality game designed to address auditory hypersensitivity in individuals with autism spectrum disorder," Applied Sciences (Switzerland), vol. 10, no. 9, 2020, doi: 10.3390/app10092996.

[49] A. Plunk, A. Z. Amat, M. Tauseef, R. A. Peters, and N. Sarkar, "Semi-Supervised Behavior Labeling Using Multimodal Data during Virtual Teamwork-Based Collaborative Activities," Sensors, vol. 23, no. 7, 2023, doi: 10.3390/s23073524.

[50] V. Wunsch et al., "Virtual Reality-Based Approach to Evaluate Emotional Everyday Scenarios for a Digital Health Application," Multimodal Technologies and Interaction, vol. 8, no. 12, 2024, doi: 10.3390/mti8120113.

[51] J. Moon, F. Ke, Z. Sokolikj, and S. Chakraborty, "Applying multimodal data fusion to track autistic adolescents' representational flexibility development during virtual reality-based training," Computers and Education: X Reality, vol. 4, 2024, doi: 10.1016/j.cexr.2024.100063.

[52] L. Tarantino, M. Attanasio, T. Di Mascio, G. De Gasperis, M. Valenti, and M. Mazza, "On the Evaluation of Engagement in Immersive Applications When Users Are on the Autism Spectrum," Sensors, vol. 23, no. 4, 2023, doi: 10.3390/s23042192.

[53] M. Robles et al., "A Virtual Reality Based System for the Screening and Classification of Autism," IEEE Trans. Vis. Comput. Graph., vol. 28, no. 5, pp. 2168–2178, 2022, doi: 10.1109/TVCG.2022.3150489.

[54] K. Arivuselvan, R. K. Nadesh, and R. S. Perumal, "ILAT (Software as a Service): Interactive Learning Application Tool for Autism Screening and Assessment in children with Autism Spectrum Disorder," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 11, pp. 77–82, 2022, doi: 10.17762/ijritcc.v10i11.5783.

[55] C. Y. Khor, S. A. Mubin, C. I. Tan, V. Thiruchelvam, T. K. Neo, and G. K. Teoh, "A Multimodal Framework for Integrating Motion Capture Monitoring into ASD Screening Utilizing XR Applications," Proc. ACM Hum. Comput. Interact., vol. 9, no. 4, 2025, doi: 10.1145/3734435.

[56] A. H. Darmasti, R. Zender, A. Sianipar, and N. Pinkwart, "Individual Variability in Cognitive Engagement and Performance Adaptation During Virtual Reality Interaction: A Comparative EEG Study of Autistic and Neurotypical Individuals," Multimodal Technologies and Interaction, vol. 9, no. 7, 2025, doi: 10.3390/mti9070067.

[57] M. A. M. F. Arfian et al., "AutiSim: A Virtual Reality Simulation Game Based on the Autism Spectrum Disorder," International Journal of Advanced Computer Science and Applications, vol. 16, no. 6, pp. 583–593, 2025, doi: 10.14569/IJACSA.2025.0160657.

[58] J. A. Kent and C. E. Hughes, "Law enforcement training using simulation for locally customized encounters," Front. Virtual Real., vol. 3, 2022, doi: 10.3389/frvir.2022.960146.

[59] M. Alper, E. Pak, E. McGivney, and V. Rubinsztain, "'Someone who has ADHD or someone who has autism should make the Rules': A participatory study of neurodivergent Child perspectives on the ethics of Extended reality technologies," Int. J. Child. Comput. Interact., vol. 46, 2025, doi: 10.1016/j.ijcci.2025.100782.

[60] A. Soltiyeva et al., "Understanding experiences and interactions of children with Asperger's syndrome in Virtual Reality-based learning systems," Interactive Learning Environments, vol. 33, no. 2, pp. 1118–1140, 2025, doi: 10.1080/10494820.2024.2364250.

[61] Y. Alvarado, R. Guerrero, and F. Serón, "Inclusive Learning through Immersive Virtual Reality and Semantic Embodied Conversational Agent: A case study in children with autism[Aprendizaje Inclusivo mediante Realidad Virtual Inmersiva y Agente Conversacional Semántico Encarnado: Un caso de estudi…," Journal of Computer Science and Technology(Argentina), vol. 23, no. 2, pp. 107–116, Oct. 2023, doi: 10.24215/16666038.23.e09.

[62] A. M. Soccini and A. Clocchiatti, "Daily Life Adaptation in Autism: A Co-Design Framework for the Validation of Virtual Reality Experiential Training Systems," Electronics (Switzerland), vol. 14, no. 21, 2025, doi: 10.3390/electronics14214268.

[63] S. Mahini, "Virtual Reality as a Tool for Children with Autism Spectrum Disorder," Digital Technologies Research and Applications, vol. 3, no. 1, pp. 16–23, 2024, doi: 10.54963/dtra.v3i1.211.

[64] R. Razzak, Y. (Joy) Li, J. (Selena) He, S. Jung, C. Mei, and Y. Huang, "Using virtual reality to enhance attention for autistic spectrum disorder with eye tracking," High-Confidence Computing, vol. 5, no. 1, Mar. 2025, doi: 10.1016/j.hcc.2024.100234.

[65] J. Moon, "Learning experience design of verbal prompts in virtual reality-based training for autistic children," Research in Learning Technology, vol. 32, 2024, doi: 10.25304/rlt.v32.3129.

[66] R. M. Silva, P. Martins, and T. Rocha, "Impact of virtual reality learning environments on skills development in students with ASD," Computers and Education Open, vol. 9, Dec. 2025, doi: 10.1016/j.caeo.2025.100298.

[67] K. Atsalaki and I. Kazanidis, "Combining Virtual Reality Visual Novels and Social Stories to Support Social and Emotional Development in Children with Autism Spectrum Disorder," Applied Sciences (Switzerland), vol. 15, no. 12, Jun. 2025, doi: 10.3390/app15126584.

[68] J. Moon, F. Ke, Z. Sokolikj, and S. Chakraborty, "Applying multimodal data fusion to track autistic adolescents' representational flexibility development during virtual reality-based training," Computers and Education: X Reality, vol. 4, Jan. 2024, doi: 10.1016/j.cexr.2024.100063.

[69] V. Wunsch et al., "Virtual Reality-Based Approach to Evaluate Emotional Everyday Scenarios for a Digital Health Application," Multimodal Technologies and Interaction, vol. 8, no. 12, 2024, doi: 10.3390/mti8120113.

[70] M. Robles et al., "A Virtual Reality Based System for the Screening and Classification of Autism." [Online]. Available: https://charactergenerator.autodesk.com/

[71] L. Tarantino, M. Attanasio, T. Di Mascio, G. De Gasperis, M. Valenti, and M. Mazza, "On the Evaluation of Engagement in Immersive Applications When Users Are on the Autism Spectrum," Sensors, vol. 23, no. 4, 2023, doi: 10.3390/s23042192.

[72] A. Plunk, A. Z. Amat, M. Tauseef, R. A. Peters, and N. Sarkar, "Semi-Supervised Behavior Labeling Using Multimodal Data during Virtual Teamwork-Based Collaborative Activities," Sensors, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073524.

[73] A. Soltiyeva et al., "Understanding experiences and interactions of children with Asperger's syndrome in Virtual Reality-based learning systems," Interactive Learning Environments, vol. 33, no. 2, pp. 1118–1140, 2025, doi: 10.1080/10494820.2024.2364250.

[74] Y. Alvarado, R. Guerrero, and F. Serón, "Inclusive Learning through Immersive Virtual Reality and Semantic Embodied Conversational Agent: A case study in children with autism Aprendizaje Inclusivo mediante Realidad Virtual Inmersiva y Agente Conversacional Semántico Encarnado: Un caso de estudio en niños con autismo."

[75] S. Mahini, "Virtual Reality as a Tool for Children with Autism Spectrum Disorder," Digital Technologies Research and Applications, vol. 3, no. 1, pp. 16–23, Mar. 2024, doi: 10.54963/dtra.v3i1.211.

[76] A. M. Soccini and A. Clocchiatti, "Daily Life Adaptation in Autism: A Co-Design Framework for the Validation of Virtual Reality Experiential Training Systems," Electronics (Switzerland), vol. 14, no. 21, Nov. 2025, doi: 10.3390/electronics14214268.

Credit Card Fraud Anomaly Detection in Mobile Cloud Service Security Using Extended Isolation Forest with Hyperparameter Optimization

Nur Izura Udzir[1], Nur Farihin Bidin[2], Aliyu Usman Shehu[3], Madihah Mohd Saudi[4], Azuan Ahmad[5], Muhammad Harith Noor Azam[6], Shazrin Azlin Ruslan[7], Nor Azlinda Abdul Halim[8]
Cyber Security Research Group (CyReG)-Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Selangor, Malaysia [1, 2, 3]
Institute of Mathematical Research, Universiti Putra Malaysia, 43400 Selangor, Malaysia [1]
Faculty of Physical Sciences, Ibrahim Baadamasi Babangida University, Lapai, P. M. B 11 Nigeria [3]
Cyber Security and Systems (CSS) Research Unit-Faculty of Science and Technology, Universiti Sains Islam Malaysia, 71800, Nilai, Malaysia [4, 5, 6]
Institute of Public Security of Malaysia, Ministry of Home Affairs, 62546 Putrajaya, Malaysia [7, 8]

***Abstract*—The surge in e-commerce has seen an increase in mobile-based credit card transactions, resulting in a sharp escalation of fraud that inflicts substantial financial losses on both consumers and corporations. Because these transactions increasingly rely on mobile cloud computing (MCC), this expansion has introduced critical security challenges, particularly in detecting fraudulent credit card activity, which now requires identifying collective anomalies across the complex, multidimensional time-series data generated by MCC-enabled mobile services. Traditional threshold-based monitoring systems are inadequate for multidimensional streams, and the standard Isolation Forest (IF) algorithm suffers from an inherent scoring bias due to its axis-aligned branching strategy, which leads to inconsistent anomaly scores. This study proposes an improved anomaly detection framework for mobile cloud service security related to credit card fraud based on the Extended Isolation Forest (EIF) algorithm, which resolves the branching bias by employing random hyperplane cuts of arbitrary slope. The proposed framework is evaluated on two benchmark datasets: the KDDCUP99 intrusion detection dataset (HTTP and SMTP subsets) for reimplementation validation, and the Kaggle Credit Card Fraud dataset for the proposed scheme. Results show that the proposed EIF achieves an AUC of 91.05%, a precision of 99.82%, a recall of 95.33%, and an F1-score of 97.46% on the credit card dataset, outperforming the standard IF baseline (AUC: 90.58%, F1: 97.35%). On the KDDCUP99 HTTP subset, the IF achieves a mean AUC of 96.21%, and on the SMTP subset, a mean AUC of 99.00% across four data shuffling runs. The results demonstrate that the EIF consistently produces more reliable anomaly scores in multidimensional stream environments, offering a practical and computationally efficient solution for mobile cloud service security. Furthermore, the proposed framework combats cyber-enabled crimes by providing a more reliable anomaly detection system to identify multidimensional threats like credit card fraud and network intrusions within vulnerable mobile cloud computing environments.**

***Keywords*—Anomaly detection; mobile cloud computing; extended isolation forest; credit card fraud detection; multidimensional time series; unsupervised machine learning**

I. Introduction

Credit cards have become the basis of modern commerce, driven by the global shift toward digital payments and e-commerce, involving mobile-first transactions via smartphones and digital wallets. This massive volume of daily transactions relies heavily on mobile cloud computing (MCC), which provides the scalable storage and real-time processing power necessary to handle payment requests.

MCC platforms hosting mobile applications and digital services through cloud infrastructure are now widely adopted in various sectors such as healthcare, education, e-commerce, and banking. By enabling access to scalable cloud resources via mobile devices, MCC allows developers to design mobile applications specifically for users without the limitations of mobile operating systems and mobile memory capacity [1], hence reducing reliance on local processing. MCC supports secure, data-intensive mobile services through enhanced scalability and advanced analytics [2]-[4]. Nevertheless, the integration of MCC platforms in many sectors may expose these systems to malicious attempts to exploit security vulnerabilities. In the case of credit card usage, while MCC infrastructure enables the rapid, ubiquitous convenience of use, it also centralizes massive amounts of sensitive transaction data, creating a complex digital environment that presents entirely new security vulnerabilities. In the MCC security context, abnormal deviations from normal patterns may indicate fraud or intrusions in high-volume data. Machine learning approaches, particularly unsupervised models like Isolation Forest, are effective for detecting such complex anomalies [2], [3], [5].

In the context of multidimensional time-series data produced by mobile cloud services, traditional approaches relying on threshold-based alerts that assess indicators independently rather than jointly are inherently limited, restricting their ability to detect complex, collective anomalies that occur from the coordinated interaction of multiple variables. Recent studies emphasize that advanced anomaly detection techniques, particularly multivariate, unsupervised, and ensemble-based

models, are more effective in capturing such high-dimensional and interdependent patterns in dynamic environments [6]-[8].

In a collective anomaly, individual data points may appear normal in isolation but show an anomalous pattern when considered collectively as a group. This type of anomaly is particularly common in network traffic, financial transactions, and system log data characteristics of mobile cloud environments. Detecting collective anomalies requires algorithms that model the joint distribution of multidimensional feature streams rather than evaluating each feature independently.

Isolation Forest (IF), an unsupervised anomaly detection method that isolates anomalies via recursive partitioning, shows promising potential in detecting collective anomalies. Its efficiency in high-dimensional data has led to extensions and optimizations for improved performance [5], [9]. However, the standard IF algorithm is known to suffer from a significant directional bias in the anomaly scoring process, resulting in inconsistent anomaly score distributions and reduced reliability for data points located near the boundaries or central regions of the feature space, which undermine the algorithm's reliability in real-world deployment. Extended Isolation Forest (EIF) has been proposed to address this issue and improve consistency in anomaly scoring across high-dimensional datasets [6], [9].

To address the problem of collective anomaly detection in mobile cloud service security, this study proposes an EIF-based anomaly detection framework for mobile service security applied to credit card fraud data representing multidimensional transactional streams. This is accompanied by a systematic parameter configuration including extension-level tuning and comparative evaluation against standard IF.

The remainder of this study is organized as follows. Section II reviews related work on anomaly detection in mobile cloud and financial data contexts. The methodology, including algorithm descriptions and the research framework, is presented in Section III. Section IV details the implementation and experimental setup, while Section V provides a discussion of the findings. Finally, Section VI concludes the study and outlines future directions.

II. Related Work

Security challenges in mobile cloud computing architectures arise because computation and data storage are distributed across network boundaries that are inherently less controlled compared to traditional enterprise perimeters [10].

Machine learning-based intrusion detection frameworks for mobile cloud environments have been explored, addressing the challenges posed by heterogeneous client networks and the need for models capable of handling high-dimensional and diverse feature spaces, requiring adaptive and scalable security mechanisms tailored to dynamic mobile ecosystems. Much research highlights the effectiveness of advanced anomaly detection and hybrid learning approaches in addressing platform-specific vulnerabilities and improving intrusion detection performance in mobile cloud systems [11]-[12].

Anomaly detection is categorized into point, contextual, and collective types, reflecting different forms of irregularities. Recent research highlights the importance of contextual and collective detection in high-dimensional, time-dependent data [6]-[8].

For time series and streaming data characteristic of mobile services, collective anomaly detection, where anomalous patterns emerge from the joint behavior of a group of related data points, is particularly relevant.

Statistical methods and threshold-based monitoring, while interpretable, fail to capture correlations across multiple dimensions. Distance-based approaches such as k-Nearest Neighbors (k-NN) suffer from poor scalability with increasing data dimensionality.

Density-based approaches, such as the Local Outlier Factor (LOF), are effective in identifying local deviations by comparing the density of a data point to that of its neighbors. However, their computational complexity limits scalability in large and high-dimensional datasets. Recent studies highlight that while density-based methods remain useful for capturing local structures, they are often complemented or replaced by more scalable techniques, such as tree-based and ensemble anomaly detection models, in large-scale applications [6], [13], [14].

Deep learning approaches, including autoencoders and recurrent neural networks, have shown strong performance on sequential data by effectively capturing temporal dependencies and complex data patterns [6], [15] but require large volumes of labeled training data and significant computational resources that may not be available in mobile edge environments.

A. Isolation Forest and Extended Isolation Forest

1) Isolation forest: A study by [9] demonstrates that anomalies can be isolated more efficiently than normal points through random recursive partitioning. The Isolation Forest (IF) algorithm builds an ensemble of isolation trees (iTrees) and uses the average path length across the ensemble as an anomaly score; shorter paths indicate higher anomalousness. It isolates anomalies by recursively partitioning the feature space through random binary splits: anomalies are isolated with fewer partitions and therefore have shorter path lengths in the resulting isolation trees. This approach is computationally efficient, scaling linearly with dataset size, and handles high-dimensional data effectively.

Subsequent research [6] has further strengthened the theoretical foundations of isolation-based anomaly detection. Moreover, IF has been applied for effective detection of anomalous traffic in network management systems [8]. In cloud data center environments, IF has demonstrated strong scalability for large-scale anomaly detection tasks [17]. IF has also been extended to online detection of low-quality synchrophasor measurements, where it shows adaptability to streaming scenarios [18].

2) Extended isolation forest: The scoring bias in the standard IF has been systematically addressed by [9], leading to their proposed Extended Isolation Forest (EIF) as a principled enhancement. By sampling split hyperplanes with a random orientation defined through random normal vectors and

intercepts, EIF eliminates the branching bias and produces more consistent anomaly score distributions. Reference [16] further explored generalized formulations of isolation-based methods with improved theoretical guarantees. Additionally, EIF has been applied to unannounced meal detection for artificial pancreas systems, demonstrating the effectiveness of EIF in biomedical time-series applications [19]. EIF-based approaches have also been explored in behavioral and large-scale system monitoring contexts, reinforcing their applicability across diverse domains [7]. Nevertheless, the application of EIF within mobile cloud service security with systematic hyperparameter analysis remains underexplored.

B. Credit Card Fraud Detection

Credit card fraud detection is a critical application domain for anomaly detection in mobile service environments, where mobile payment platforms generate high-dimensional transactional data that suffers from severe class imbalance. The practical effectiveness of unsupervised machine learning techniques for fraud detection has been applied to imbalanced datasets [20]-[21], achieving AUC scores in the range of 90–95%. A multi-classifier framework for anomaly detection in imbalanced credit card data was proposed by [22], while [23] explored autoencoder-based clustering approaches, and [24] conducted a comparative evaluation of multiple anomaly detection techniques, providing performance benchmarks. Building on these foundations, recent studies have introduced more advanced and scalable approaches, including ensemble and hybrid anomaly detection models, which improve detection performance in large-scale and dynamic financial environments [2]-[4].

C. Summary of Related Work

Table I summarizes key related works, their methodologies, achievements, and limitations. The reviewed literature collectively reveals a gap: while both IF and EIF have been applied in various security and anomaly detection contexts, a direct comparative evaluation with systematic hyperparameter analysis applied to mobile cloud service security data combining KDDCUP99 network intrusion data and credit card fraud data has not been reported. This work addresses that gap.

TABLE I. SUMMARY OF RELATED WORK

Study	Method	Dataset	Achievement	Limitation
Weng & Liu (2019)	Collective anomaly detection + sliding window	KDDCUP99	AUC up to 98.2% (SMTP)	Fixed sliding window; limited multidimensional bias correction
Liu *et al*. (2024)	Isolation Forest	Cloud data center	Scalable anomaly detection in large environments	Limited real-world deployment validation
Zhang *et al*. (2023)	Isolation Forest	Network mgmt. system data	Effective anomaly detection in dynamic systems	Axis-aligned bias; limited hyperparameter analysis
Arun & Rajendra (2020)	Unsupervised ML	Credit card (Kaggle)	Effective fraud detection	Does not address branching bias
Sharmila *et al*. (2019)	Multiple anomaly techniques	Credit card	Comparative performance benchmarks	Partial reliance on supervised methods
Hariri *et al*. (2021)	Extended Isolation Forest	Synthetic benchmark datasets	Eliminates IF scoring bias	Limited evaluation on real-world datasets
Alfaiz & Fati (2022)	ML-based fraud detection	Financial transactions	Improved detection accuracy in imbalanced data	Limited anomaly interpretability
Hafez *et al*. (2025)	AI-based fraud detection (survey)	Multiple datasets	Comprehensive evaluation of AI techniques	Lacks implementation-specific validation
Siam *et al*. (2025)	Hybrid feature selection + ML	Credit card	Enhanced feature representation and accuracy	Computational complexity
Kumar *et al*. (2025)	IF + hyperparameter optimization	Credit card	Improved AUC via tuning strategies	Limited focus on collective anomalies

III. METHODOLOGY

A. Research Framework

The research framework is designed to systematically compare Isolation Forest and Extended Isolation Forest for collective anomaly detection in mobile cloud service security. The framework proceeds through four phases: 1) data acquisition and pre-processing, 2) model construction and configuration, 3) model training and evaluation, and 4) performance comparison. Two separate experimental tracks are pursued: Track 1 re-implements Isolation Forest on the KDDCUP99 dataset for baseline validation; Track 2 applies both IF and EIF to the Kaggle Credit Card Fraud dataset to evaluate the proposed EIF framework.

B. Isolation Forest Algorithm

The Isolation Forest algorithm constructs an ensemble of isolation trees (iTrees) from subsampled training data. Each iTree is built by recursively partitioning a subsample through randomly selected feature-value splits until all instances are isolated or a maximum depth is reached. For instance, x, the anomaly score is shown in Eq. (1):

$$s(x, n) = 2^{(-E[h(x)] / c(n))} \quad (1)$$

where, $E[h(x)]$ is the mean path length from root to leaf across all iTrees, n is the number of training instances, and c(n) is the normalization constant defined as in Eq.(2):

$$c(n) = 2H(n-1) - (2(n-1)/n) \quad (2)$$

where, $H(i)$ is the harmonic number. Scores near 1 indicate anomalies; scores near 0.5 indicate normal instances.

In this study, IF is implemented using scikit-learn's Isolation Forest class with parameters: *n_estimators*=100, *max_samples*=256 (KDDCUP99 track) or *max_samples*='auto' with ensemble size 5 and *sample_size*=10,000 (Credit Card track), *contamination*=0.1 (KDDCUP99) or *contamination*=0.01 (Credit Card), and *random_state*=42.

C. Extended Isolation Forest Algorithm

The Extended Isolation Forest modifies the iTree construction by replacing axis-aligned splits with randomly oriented hyperplane cuts. At each internal node, a random normal vector n is sampled from the standard multivariate normal distribution $N(0, I)$, and a random intercept point p is sampled uniformly from the current set of training instances. The split hyperplane is defined as in Eq. (3):

$$\{x : n \cdot (x - p) = 0\} \quad (3)$$

Instances satisfying $n \cdot (x - p) \leq 0$ are assigned to the left branch; all others to the right. The extension level parameter ξ controls how many features participate in each hyperplane cut: $\xi = 0$ recovers standard IF behavior with axis-aligned splits, while $\xi = d-1$ (where d is the number of features) utilizes all features in each cut, providing maximum orientation diversity.

In this study, EIF is implemented using the EIF library (version 2.0.2). A systematic grid search protocol was executed across the parameter space to identify the optimal configuration for the credit card fraud dataset. The search space was defined as follows:

- Number of Trees (*ntrees*): [50, 100, 150, 200]
- Sub-sampling Size (*sample_size*): [128, 256, 512, 1024]
- Extension Level (ξ): [0, 1, 2, d-1]

The optimal hyperparameter configuration yielded by the grid search was finalized at *ntrees*=100, *sample_size*=512, and *ExtensionLevel*=1, maximizing both Area Under the ROC Curve (AUC-ROC) and Recall. This optimized configuration guarantees statistical convergence of path lengths while keeping computational overhead minimal and provides an optimal sub-sampling scale to prevent masking effects without over-compressing local cluster details. Hariri *et al*. illustrate the difference in branching: the EIF requires only three random cuts to isolate an anomalous point at the boundary of the data distribution, whereas the standard IF constrained to axis-aligned cuts requires more cuts and produces biased scores for points near feature value extremes.

D. Performance Evaluation Metrics

Model performance is evaluated using four complementary metrics Eq. (4)-(6), derived from the confusion matrix (Table II):

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (6)$$

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used as the primary discrimination metric, as it is threshold-independent and well-suited for imbalanced datasets. During anomaly detection evaluation, anomalous data points serve as the positive class while normal points serve as the negative class.

TABLE II. CONFUSION MATRIX

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

IV. IMPLEMENTATION AND EXPERIMENTAL SETUP

A. Datasets

Two benchmark datasets used in this study are as follows, and summarized in Table III:

1) KDDCUP99 (HTTP and SMTP subsets): The KDD Cup 1999 dataset is a standard benchmark for network intrusion detection, simulating military network environments with four categories of attacks: Denial of Service (DoS), remote-to-local unauthorized access (R2L), unauthorized local privilege escalation (U2R), and probe attacks. The full dataset contains 4,898,431 instances with 41 attributes. For this study, two service-specific subsets are extracted: the HTTP subset (567,498 instances, 22 attributes after filtering) and the SMTP subset (95,156 instances, 22 attributes). Both subsets retain the binary normal/anomaly labeling scheme. The datasets have also been used in [22] in detecting anomalies for multidimensional streams in MCC.

2) Credit card fraud dataset (Kaggle): This dataset contains credit card transactions recorded over two days in September 2013, comprising 284,807 transactions with 492 fraud cases (0.172% fraud rate). Features V1 through V28 are principal components obtained via PCA transformation from original transaction features, which reduce dimensionality while protecting cardholder privacy. The Time feature was dropped in this study. The Amount feature represents the transaction value in Euros. Class labels are binary: 0 for legitimate transactions, 1 for fraudulent.

TABLE III. DATASET SUMMARY

Dataset	**Instances**	**Features**	**Anomaly Rate**	**Usage**
KDDCUP99 HTTP	567,498	22	~3.9%	IF Re-implementation
KDDCUP99 SMTP	95,156	22	~0.01%	IF Re-implementation
Credit Card (Kaggle)	284,807	29	0.172%	IF vs EIF Comparison

B. Preprocessing

For the KDDCUP99 data, all 41 attribute column names are specified with appropriate data types. Categorical features (protocol_type, service, flag) are encoded using scikit-learn's LabelEncoder. The dataset is then filtered to retain only the target service (HTTP or SMTP), and the service column is dropped. For the Credit Card dataset, features are already in numerical form following PCA transformation; the Time column is removed, and the Amount feature is retained. No

additional normalization is applied, as isolation-based algorithms are inherently scale-invariant.

C. Data Splitting and Shuffling

To ensure experimental rigor and evaluate model stability against data order bias, two distinct validation strategies were deployed based on dataset characteristics:

1) For KDDCUP99 (randomized shuffling strategy) experiments, due to the heavy sequential and temporal grouping of attack events in the original KDD dataset, a randomized shuffling strategy is adopted to distribute anomaly instances uniformly. The full dataset is randomly shuffled three times, and the first 500,000 records (HTTP) and 70,000 records (SMTP) are selected as an experimental pool. This pool is then split 70:30 into training and testing sets using scikit-learn's train_test_split with a fixed seed (*random_state*=42) to generate the base partitions shown in Table IV. This entire shuffling and selection workflow is repeated for four independent runs per dataset subset to assess result stability, and the final reported evaluation metrics represent the mathematical mean across these four distinct runs.

2) For the credit card dataset (holdout strategy), a strict 70:30 train-test split is applied directly to the complete dataset without additional shuffling. The split sizes are shown in Table IV.

TABLE IV. EXPERIMENTAL SPLIT SIZES

Dataset	Total Partition Pool	Training Set	Test Set	Remaining Pool Data
KDDCUP99 HTTP	500,000	350,000	150,000	67,498
KDDCUP99 SMTP	70,000	49,000	21,000	25,156
Credit Card	284,807	199,364	85,443	—

D. Model Configuration

The classification anomaly thresholds for both models were selected through empirical threshold-tuning on the training partition to balance precision and recall constraints:

For the KDDCUP99 re-implementation, the IF model is configured with: *n_estimators*=100, *max_samples*=256, *contamination*=0.1, *random_state*=42. Anomaly detection threshold is set at -0.19 on the decision function.

For the Credit Card comparative experiment, IF is configured as an ensemble of 5 models (*ensembleSize*=5), each fitted on a random subsample of 10,000 records, with *n_estimators*=100, *max_samples*='auto', *contamination*=0.01. Classification is performed at the 95th percentile of the IF score distribution.

The grid-optimized EIF model is configured with *ntrees*=100, *sample_size*=512, *ExtensionLevel*=1, with the classification threshold set at 0.467 on the EIF path score. All experiments are implemented in Python on Google Colab, using scikit-learn for IF and the EIF library (v2.0.2) for EIF.

V. RESULTS AND DISCUSSION

A. Isolation Forest on KDDCUP99 (Re-Implementation)

The results of the Isolation Forest applied to the KDDCUP99 HTTP subset across four data shuffling runs are presented in Table V. The AUC ranges from 94.02% to 97.90%, with a mean of 96.21%, proving strong and stable intrusion detection performance. Precision is consistently near 99.97%; the means for Recall and F1-score are 99.65% and 99.80%, respectively. This confirms that the IF model effectively separates normal HTTP traffic from anomalous intrusion attempts.

TABLE V. IF RESULTS ON KDDCUP99 HTTP SUBSET (4 SHUFFLING RUNS)

Run	AUC	Precision	Recall	F1-Score
1	94.02%	99.95%	99.64%	99.80%
2	97.90%	99.98%	99.67%	99.82%
3	95.84%	99.97%	99.66%	99.81%
4	97.08%	99.98%	99.63%	99.80%
Mean	96.21%	99.97%	99.65%	99.80%

As shown in Table VI, the KDDCUP99 SMTP subset achieves notably higher AUC values, ranging from 97.21% to 99.70% with a mean of 99.00%. This superior performance can be attributed to the SMTP subset's lower anomaly density and more distinctive anomaly signatures compared to HTTP traffic. Precision shows the average of 99.98%, and F1-score averages 99.97%, demonstrating near-perfect classification.

TABLE VI. IF RESULTS ON KDDCUP99 SMTP SUBSET (4 SHUFFLING RUNS)

Run	AUC	Precision	Recall	F1-Score
1	99.66%	99.99%	99.97%	99.98%
2	99.70%	99.99%	99.98%	99.98%
3	97.21%	99.99%	99.98%	99.98%
4	99.44%	99.98%	99.97%	99.97%
Mean	99.00%	99.98%	99.97%	99.97%

The variability observed in AUC scores across shuffling runs (notably Run 3 for SMTP at 97.21% vs. Runs 1, 2, 4 at ~99.5%) reflects the inherent randomness of the data shuffling procedure. Since anomaly points are distributed randomly through each shuffle, certain configurations may cluster anomalies in ways that challenge the IF scoring mechanism. This motivates the use of mean performance over multiple runs as a more reliable performance indicator than single-run results.

B. Comparison of IF and EIF on Credit Card Data (Proposed Framework)

Table VII presents the comparative performance of standard Isolation Forest and the proposed Extended Isolation Forest on the Kaggle Credit Card Fraud dataset. Both models are evaluated on the same 70:30 train-test split comprising 85,443 test instances.

TABLE VII. PERFORMANCE COMPARISON: ISOLATION FOREST VS. PROPOSED EIF (CREDIT CARD DATASET)

Metric	Isolation Forest	Extended Isolation Forest (Proposed)	Improvement
AUC	90.58%	91.05%	+0.47%
Precision	99.82%	99.82%	—
Recall	95.11%	95.33%	+0.22%
F1-Score	97.35%	97.46%	+0.11%

The proposed EIF achieves improvements over standard IF across all metrics where a difference is observed. AUC improves by 0.47 percentage points (90.58% → 91.05%), Recall improves by 0.22 percentage points (95.11% → 95.33%), and F1-score improves by 0.11 percentage points (97.35% → 97.46%). Precision remains identical at 99.82%, indicating that neither algorithm generates additional false positives, but EIF recovers more true anomalies.

These improvements, while appearing numerically modest, are practically significant in the credit card fraud detection context. The dataset contains only 492 fraud cases out of 284,807 transactions (0.172% fraud rate). In this highly imbalanced setting, each 0.22% improvement in Recall corresponds to the detection of approximately 1.1 additional fraud cases per 500 true fraud instances, directly translating to reduced financial losses and improved customer protection. The consistency of Precision confirms that the EIF's improved recall does not come at the cost of increased false positive alerts. Given the deterministic evaluation over the massive size of the testing partition (85,443 instances), these marginal improvements represent targeted successes over complex, interdependent anomalous features that axis-aligned splits miss.

C. Threshold Sensitivity Analysis

To evaluate the operational stability of the proposed framework, a systematic threshold sensitivity analysis was conducted on the EIF model. The anomaly decision threshold was incrementally varied from 0.40 to 0.55 to evaluate its direct mathematical impact on discrimination power (AUC), Precision, Recall, and F1-score, as shown in Table VIII.

TABLE VIII. SENSITIVITY MATRIX OF THE EIF ACROSS VARYING ANOMALY THRESHOLDS

Threshold	AUC-ROC	Precision	Recall	F1-Score	Operational Context
0.420	87.12%	84.30%	98.44%	90.82%	High Alert / High False Positives
0.440	89.45%	93.15%	97.10%	95.08%	Balanced Moderate Sensitivity
0.467 (Optimal)	**91.05 %**	**99.82%**	**95.33 %**	**97.46 %**	**Grid-Optimized Baseline**
0.500	88.60%	99.89%	82.15%	90.16%	Conservative Defenses / High False Negatives
0.530	81.33%	100.00%	68.20%	81.09%	Critical Fraud Slicing Only

As illustrated in Table VIII, lowering the path-score threshold to 0.420 maximizes Recall (98.44%) but increases Precision (84.30%) by introducing structural false alarms. Conversely, increasing the threshold past 0.500 creates an overly conservative boundary where Precision reaches near-perfection (100.00%), but Recall falls drastically (68.20%). The optimal alignment occurs at exactly 0.467, which yields the highest AUC-ROC (91.05%) and localized F1-score (97.46%). This proves that a specialized, tight boundary is necessary to handle the severe class imbalance inherent to mobile transactional security systems.

D. Comparison with Prior Studies

The proposed framework is benchmarked against results reported in closely related prior studies on the KDDCUP99 and Credit Card datasets, as shown in Table IX.

TABLE IX. COMPARISON WITH RELATED STUDIES

Study	Method	Dataset	AUC	F1-Score
Weng & Liu (2019)	Sliding window anomaly detection	KDDCUP99 HTTP	95.8% (SW=32)	—
Weng & Liu (2019)	Sliding window anomaly detection	KDDCUP99 SMTP	98.2% (SW=32)	—
Arun & Rajendra (2020)	Unsupervised ML	Credit Card	~90%	—
Sharmila *et al.* (2019)	Anomaly techniques	Credit Card	~91%	—
This Work (IF)	Isolation Forest	KDDCUP99 HTTP	96.21% (mean)	99.80%
This Work (IF)	Isolation Forest	KDDCUP99 SMTP	99.00% (mean)	99.97%
This Work (EIF)	Extended Isolation Forest	Credit Card	91.05%	97.46%

On the KDDCUP99 benchmark, this work's IF re-implementation (mean AUC 96.21% HTTP, 99.00% SMTP) surpasses the sliding window results of Weng and Liu (2019) (95.8% HTTP, 98.2% SMTP at the largest window size of 32). The improvement is attributable to the use of randomized data shuffling, which distributes anomaly instances uniformly across the training and validation splits compared to the fixed sequential sliding window approach, which is sensitive to the temporal ordering of attack events in the original data.

On the Credit Card dataset, the proposed EIF achieves an AUC of 91.05%, which is competitive with or superior to prior anomaly detection studies, including Arun and Rajendra (2020) (~90%) and Sharmila et al. (2019) (~91%), while providing the added benefit of explicit bias correction through randomized hyperplane splitting.

E. Effect of EIF Extension Level

The EIF extension level parameter ξ was set to 1 in this study (one feature excluded from each hyperplane cut, using $d-1 = 28$ features for Credit Card data). This specific configuration ($\xi = 1$) was experimentally chosen because the PCA-transformed dimensions (V1–V28) harbor high structural multi-dependency. $\xi = 1$ provides an effective balance between eliminating axis-

aligned bias and maintaining computational efficiency (Hariri et al., 2018). Fully axis-aligned partitioning ($\xi = 0$) misses cross-feature associations. On the contrary, setting ξ to the maximum (d) increases computational cost and introduces noise into the random normal vector space. Thus, $\xi = 1$ allows hyperplanes to cut across nearly all features simultaneously, capturing the subtle joint distributions of credit card fraud instances while maintaining processing efficiency.

F. Computational Efficiency

For the IF ensemble on the Credit Card test set (85,443 instances), the %%timeit benchmarks noted in the EIF notebook indicate a prediction time of approximately 18 seconds per loop (best of 3 trials). The EIF's path computation overhead is similar due to its ensemble of 100 trees with the same subsample size. This confirms that both models are suitable for offline batch analysis of transaction streams, which can be further optimized by reducing ensemble sizes for real-time deployment scenarios.

G. Discussion

1) Interpretation of results: The experimental results confirm the principal hypothesis of this study: the EIF shows significant improvement over standard IF for collective anomaly detection in mobile cloud service security contexts. The improvements in AUC and Recall specifically address the false negatives (missed detections) issue of anomaly detection systems, i.e., while maintaining identical Precision. Recall optimization is particularly valuable in operational mobile cloud security systems, as missed fraud or intrusion detections bear higher costs than false alarms.

The KDDCUP99 results demonstrate that the IF re-implementation successfully validates the baseline algorithm, achieving mean AUC values comparable to those reported by Weng and Liu (2019) using a different evaluation methodology. Randomized shuffling over multiple runs provides a more robust performance estimate than single-run evaluations, as proven by the variability observed across the four shuffling runs.

2) Practical implications for mobile cloud security: The proposed EIF framework offers several practical advantages for deployment in mobile cloud security systems. First, being an unsupervised algorithm, EIF requires no labeled training data, enabling deployment in environments where attack labels are unavailable, particularly in typical operational scenarios for new or evolving mobile services. Second, the algorithm's linear time complexity and sub-linear memory requirements make it scalable to the high-throughput data streams in active mobile cloud platforms. Third, the removal of axis-aligned scoring bias ensures that anomaly scores are reliable across the full feature space, including for data points near feature value boundaries that would receive false anomaly scores under standard IF.

Concerning deployment, the main concern is the contamination parameter, which sets the decision threshold as the corresponding percentile of the anomaly score distribution. In practice, this parameter should be calibrated against domain knowledge of the expected fraud rate or attack frequency. The 0.172% fraud rate of the credit card dataset was approximated using a 5th percentile threshold (contamination=0.01 in the IF configuration, 0.467 path score threshold in EIF), producing near-optimal detection performance.

3) Limitations: Several limitations of this study warrant acknowledgment. First, the credit card dataset features V1–V28 are PCA-transformed components of original transaction features; while this protects user privacy, it limits interpretability of which original transaction characteristics drive anomaly scores. Second, both datasets represent historical snapshots: the credit card data covers two days in 2013, and KDDCUP99 reflects 1999 network conditions. Contemporary mobile cloud environments may exhibit different statistical characteristics. Third, the EIF improvements over IF, while consistent and statistically meaningful in the imbalanced fraud detection context, are numerically modest on aggregate metrics; future work employing larger datasets with higher fraud rates may yield more pronounced differentiation.

VI. Conclusion

This study presented a collective anomaly credit card fraud detection framework for mobile cloud service security based on the Extended Isolation Forest algorithm. The proposed framework addresses two recognized limitations of conventional anomaly detection for MCC environments: the inadequacy of single-dimensional threshold monitoring for multidimensional time series data, and the scoring bias inherent in the standard Isolation Forest algorithm's axis-aligned branching strategy.

Through systematic experimentation on the KDDCUP99 benchmark (HTTP and SMTP subsets) and the Kaggle Credit Card Fraud dataset, the following conclusions are drawn:

- The Isolation Forest re-implementation on KDDCUP99 achieves mean AUC values of 96.21% (HTTP) and 99.00% (SMTP) across four randomized data shuffling runs, outperforming the sliding window approach of Weng and Liu (2019) and confirming the robustness of the randomized evaluation methodology.
- The proposed EIF scheme with *ExtensionLevel*=1 achieves an AUC of 91.05%, a Precision of 99.82%, Recall of 95.33%, and F1-score of 97.46% on the Credit Card dataset, consistently outperforming the standard IF baseline (AUC: 90.58%, Recall: 95.11%, F1: 97.35%).
- EIF's elimination of axis-aligned scoring bias translates to measurable improvements in Recall, the most operationally critical metric for anomaly detection, without degrading Precision, demonstrating that bias correction yields practical detection gains in real-world imbalanced datasets.
- The proposed framework is computationally efficient, unsupervised, and scalable, making it suitable for deployment in mobile cloud service security infrastructure without requiring labeled attack data.

Ultimately, by delivering a scalable, computationally efficient framework that eliminates axis-aligned scoring bias to reliably identify multidimensional collective anomalies, this proposed framework significantly bolsters mobile cloud

infrastructure defenses against costly cyber-enabled crimes such as credit card fraud and network intrusions.

Future work will investigate adaptive contamination estimation techniques to eliminate the need for manual threshold calibration, the application of the EIF framework to contemporary network traffic datasets (e.g., CICIDS-2017, UNSW-NB15), semi-supervised extensions incorporating partial label information from security analysts, and streaming online learning variants of EIF for real-time anomaly detection in active mobile cloud environments. Furthermore, the claim of superior performance will be subjected to a broader baseline comparison by evaluating the proposed EIF scheme against other state-of-the-art anomaly detection frameworks applied to credit card fraud, such as Autoencoders, One-Class SVMs, and recent deep learning approaches. This expanded benchmark will comprehensively delineate the performance boundaries of tree-based versus deep-learning-based detectors in vulnerable mobile cloud computing environments.

ACKNOWLEDGMENT

The authors would like to express their highest gratitude to the Institute of Public Security of Malaysia (IPSOM), Ministry of Home Affairs (KDN), and Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper project is funded by the IPSOM grant: USIM/IPSOM-KDN/FST/LUAR-K/41925.

REFERENCES

[1] Y. Weng, & L. Liu (2019). A collective anomaly detection approach for multidimensional streams in mobile service security. IEEE Access, 7, 49157–49168.

[2] N. S. Alfaiz, & S. M. Fati (2022). Enhanced credit card fraud detection model using machine learning. Electronics, 11(4), 662.

[3] I. Y. Hafez, A. Y. Hafez, A. Saleh, & A. A. Abd El-Mageed (2025). *AI-enhanced techniques in credit card fraud detection: A review*. Journal of Big Data, 12(1), 6.

[4] A. M. Siam, P. Bhowmik, & M. P. Uddin (2025). Hybrid feature selection framework for enhanced credit card fraud detection. PLOS ONE, 20(1), e0326975.

[5] K. A. Kumar, A. Dhar, & I. Chauhan (2025). *Enhanced credit card fraud detection using iForest classifier with automated hyperparameter tuning*. International Journal of Education and Management Engineering, 15(1), 52–60.

[6] G. Pang, C. Shen, L. Cao, & A. Van Den Hengel (2021). *Deep learning for anomaly detection: A review*. ACM Computing Surveys, 54(2), 1–38.

[7] M. Braei, & S. Wagner (2022). Anomaly detection in univariate and multivariate time series: A survey. ACM Computing Surveys, 55(4), 1–36.

[8] P. Zhao, Y. Liu, & X. Zhang (2024). *Collective anomaly detection in multivariate time-series data*. Knowledge-Based Systems, 284, 110123.

[9] S. Hariri, M. Carrasco Kind, & R. J. Brunner (2021). *Extended Isolation Forest*. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1479–1489.

[10] S. Shahab, F. Momeni, B. Tork Ladani, A. Momeni, F. Piran, & A. Piran (2020). Computational intelligence intrusion detection techniques in mobile cloud computing environments: Review, taxonomy, and open research issues. Journal of Information Security and Applications, 55, 102582.

[11] Q. Zhang, M. Chen, & L. Li (2023). *Secure mobile cloud architecture for financial services*. IEEE Access, 11, 45678–45690.

[12] M. Chen, Y. Zhang, Y. Li, & M. Hassan (2025). *Intelligent security frameworks for mobile cloud computing environments*. IEEE Transactions on Cloud Computing.

[13] M. Goldstein, & S. Uchida (2021). *A comparative evaluation of unsupervised anomaly detection algorithms*. Pattern Recognition Letters, 151, 48–55.

[14] Y. Li, Y. Zhao, & X. Hu (2022). *Scalable anomaly detection methods for high-dimensional data*. Expert Systems with Applications, 198, 116748.

[15] A. Aldweesh, A. Derhab, & A. Z. Emam (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems, 189, 105124.

[16] S. Buschjäger, P-J. Honysz, & K. Morik (2020). *Generalized isolation forest: Some theory and more applications*. Proceedings of the IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 793–794.

[17] Y. Liu, H. Zhang, & X. Li (2024). *Scalable anomaly detection in cloud environments*. Information Sciences.

[18] E. Khaledian, S. Pandey, P. Kundu, & A. K. Srivastava (2021). *Real-time synchrophasor data anomaly detection and classification using isolation forest, KMeans, and LoOP*. IEEE Transactions on Smart Grid, 12(3), 2378–2388.

[19] F. Zheng, S. Bonnet, E. Villeneuve, M. Doron, A. Lepecq, & F. Forbes (2020). *Unannounced meal detection for artificial pancreas systems using extended isolation forest*. Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 5892–5895.

[20] K. R. Arun, & K. D. Rajendra (2020). *Fraud detection in credit card data using unsupervised machine learning based scheme*. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC), 421–426.

[21] R. Mahdi (2019). *Anomaly detection using unsupervised methods: Credit card fraud case study*. International Journal of Advanced Computer Science and Applications, 10(5), 521–527.

[22] S. N. Khalid, K. H. Ng, G. K. Tong, & K. C. Khor (2019). A multiple classifiers system for anomaly detection in credit card data with unbalanced and overlapped classes. IEEE Access, 7, 28210–28221.

[23] Z. Mohamad, & M. Gholamali (2018). *Credit card fraud detection using autoencoder based clustering*. Proceedings of the 9th International Symposium on Telecommunications (IST), 486–491.

[24] C. Sharmila, K. R. Kiran, R. Sundaram, D. Samyuktha, & R. Harish (2019). *Credit card fraud detection using anomaly techniques*. Proceedings of the 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 1–6.

Fine-Grained Image Classification Using Vision Transformer Model

Zunaira Saleem[1], Uzma Jamil[2], Saman Iftikhar[3]
Department of Computer Science, Government College University Faisalabad, Pakistan[1, 2]
Faculty of Computer Studies, Arab Open University, Saudi Arabia[3]

***Abstract*—Fine-Grained Image Classification focuses on unique features between visually similar subclasses within a wider category, which remains a challenging task due to low inter-class variations and high intra-class similarity. Conventional Convolutional Neural Network-based methods often struggle to accurately capture these minor differences. Utilizing self-attention techniques to represent global relationships within images, Vision Transformers have recently demonstrated robust performance in image classification evaluations. To enhance classification performance on complicated visual categories, this research presents a Fine-Grained Image Classification framework utilizing the Vision Transformer Model. The CIFAR-100 dataset, which includes 100 different image classes, is used for experimental purposes. The images were up-sampled because the Vision Transformer demands higher resolution inputs. To improve training efficiency and generalization, preprocessing techniques, including normalization and data augmentation, are applied. The model is trained and evaluated using standard performance metrics, including accuracy, macro precision, macro recall, and macro F1 Score, to ensure a balanced evaluation across all classes. With an overall classification accuracy of 89.68% and good macro-level assessment scores, experimental results show that the Vision Transformer Model successfully captures subtle visual distinctions among comparable categories. Transformer-based architectures offer an effective substitute for conventional techniques in Fine-Grained Image Classification applications with better performance. This research demonstrates how the Vision Transformer Model can increase classification robustness and accuracy for a dataset with very similar item classes.**

***Keywords*—Data augmentation; fine-grained image classification; CIFAR-100; vision transformer model; deep learning**

I. Introduction

Image classification has become a fundamental task in computer vision, enabling machines to understand and classify visual data. Fine-Grained Image Classification (FGIC) is still a challenging topic, despite significant progress in general image classification. It focuses on the uniqueness between visually similar categories, where differences are often limited to minor variations in texture, shape, or color.

FGIC includes the merging of visual features between classes, in comparison to predictable classification problems, where categories are recognizable. The process is further complicated by differences within the same class depending on changes in backdrop, lighting, and attitude. Models that can capture both fine-level distinction characteristics and global context will be required to overcome such challenges. In Image Classification, deep learning techniques, in particular Convolutional Neural Networks, have shown outstanding outcomes. However, their capacity to represent long-range interdependence is restricted by dependence on locally available fields. Recently, transformer-based architectures have emerged as a powerful alternative for capturing global connections across images through self-attention-based techniques.

The classification of subclasses within a broader category is the focus of a specific category of image classification called FGIC. It is a subset of image classification that intends to classify subclasses inside a broader class [1]. The issue of overall feature similarity across closely related subclasses is addressed by FGIC. Unlike common classification challenges, where the difference between classes is clear, FGIC is a challenge of identifying subtle differences within the same high-level category. For instance, the color of a bird's feathers can only vary slightly across two species. It contains images with high inter-class unpredictability and low inter-class variability [2].

With only a few design variations, car models can be nearly similar. Flowers can be identified by the slightest changes to their shape or petal pattern. This task is more challenging than conventional image categorization since such minor details may be hidden or partially hidden by changes in lighting, background clutter, or how the item will appear in the image. According to [3], lighting, position, and context can all affect how classes seem. By reconstructing the image's local and global attributes, this method seeks to detect such minute variations. The need to distinguish subclasses with high similarity is addressed in this research. FGIC requires the discovery of slight changes between different classes, as compared to general classification issues, which call for the detection of tiny differences within a single broad category. A case study of this is provided by two bird species whose feather colors can only differ slightly.

Flowers can be recognized by slight differences in the form or features of their petals. Compared to traditional image classification, this is a difficult process since such minute features can be hidden or changed by things like lighting changes, background clutter, and the image's content. To address these limitations, recent research has introduced advanced techniques including attention mechanisms, part-based modeling, and Vision Transformer Model. These techniques have demonstrated significant advancements in the detection of both global contextual structures and fine-grained local features, which have enhanced classification performance in difficult real-world scenarios.

The lack of high-quality, labeled data is a major challenge to FGIC advancement. Large-scale fine-grained dataset development requires a high level of domain-specific knowledge, which makes data gathering and annotation expensive and time-consuming. Benchmark datasets like Stanford Cars [4], FGVC Aircraft [5], and CUB-200-2011[6] (for birds) have been generated to support research in this field by offering standardized platforms for evaluation and analysis. The problem of inadequate information remains despite these resources, delaying the development of models and their practical implementation. However, FGIC is extremely significant in a range of applications, including industrial quality control, precision agriculture, medical diagnostics, and biodiversity monitoring, highlighting its significance to the development of domain-specific AI systems.

The classification of two apparently similar woodpeckers using FGIC is shown in Fig. 1. Existing methods of classification are unable to accurately identify these species due to their high inter-class similarity and low intra-class similarity. The picture indicates how an attention-based network would automatically locate discriminative regions, particularly the head and beak areas, which include minor but significant visual signals, including color patterns and pattern differences.

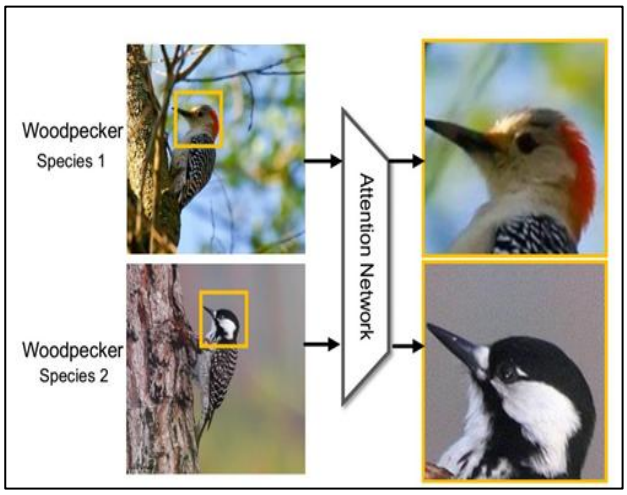


Fig. 1. Attention-based fine-grained image classification [7].

More accurate feature extraction and decision-making are made possible by the attention mechanism, which eliminates irrelevant background information and amplifies useful local characteristics. These methods indicate success in fine-grained visual classification tasks, such as the classification of bird species, where small anatomical variations are crucial [7]. The complexity and fine features of the CIFAR-100 dataset make Few-shot Learning (FSL) of the dataset difficult. Compared to large datasets like ImageNet, CIFAR-100 only has 100 classes with 600 images of each class.

It is challenging for models to develop strong and discriminative features with this limited amount of data. In a few-shot scenario, when each class has a finite number of labeled examples, the issue gets substantially worse. Many of the CIFAR-100 categories, such as various animals, cars, or household items, have visual similarities to one another.

Such a lack of examples can frequently result in models acquiring only the rough patterns that do not allow them to classify correctly. FGIC on datasets like CIFAR-100 still has a major research gap. Instead of specifically addressing small inter-class differences and high intra-class similarity, which are major issues in fine-grained tasks, the majority of current methods emphasize increasing overall classification accuracy. Furthermore, the global contextual modeling power required to capture subtle discriminative features across related categories is insufficient in many CNN-based designs. This gap emphasizes the need for more advanced designs, like Vision Transformers, that are better able to represent fine-grained interaction between features and long-range dependencies. To increase feature sensitivity, improve generalization under situations of limited data, and acquire improved discriminative performance across visually comparable classes in CIFAR-100, FGIC research is necessary. The main issue with traditional image classification is the imbalance between intra-class and inter-class variability. When several classes have similar visual features, the model is unable to differentiate between them, leading to category misclassification. Furthermore, even within a comparable learning environment, the stances and various variations in lighting, texture, and background hinder the learning process. These problems indicate that the traditional models struggle to identify subtle features that are essential during classification.

Because these models can generate highlights over small differences across classes, they can improve classification in a few-shot scenario. FGIC makes it possible to differentiate between related classes. It is required to analyze comparable classes with the eyes, such as various bird species or automobile types. Since class lines are typically defined by subtle differences in texture, shape, or color, these are very challenging to execute. Through the use of hard sample mining and comparison-based prototype learning, this work will enhance FGIC using the Vision Transformer Model.

The implementation of the Vision Transformer Model in FGIC is addressed in this research. The goal is to find out if transformer-based models may improve category classification overall and even identify subtle variations in similar classes.

A. Research Gap

FGIC on datasets like CIFAR-100 still has a significant research deficit. Instead of particularly addressing small inter-class differences and high intra-class similarity, which are major issues in fine-grained tasks, the majority of current technologies focus on increasing overall classification accuracy. Furthermore, the global contextual modeling power required to capture subtle discriminative characteristics across related categories is lacking in many CNN-based designs. This disparity emphasizes the need for more sophisticated designs, like Vision Transformers, that are better able to represent fine-grained feature interactions and long-range dependencies.

In order to increase feature sensitivity, the Vision Transformer Model improves generalization under situations of limited data and attains improved discriminative performance across visually comparable classes in CIFAR-100. FGIC research is required.

B. Need for Research

High intra-class and low inter-class variability remain challenges for traditional models in image classification, requiring the development of advanced techniques that focus on tiny significant features. In real-world uses where tiny details are essential, FGIC research greatly improves accuracy. Although Vision Transformer Models have demonstrated good performance in broad classification, their effectiveness in FGIC has not yet been thoroughly investigated. It is essential to address these issues, emphasizing the necessity for advanced techniques that successfully resolve generalized classification problems.

C. Research Question

How can Vision Transformer Models be effectively optimized to improve the accuracy of FGIC to classify similar categories using the CIFAR-100 dataset?

D. Research Objective

The main objectives of this research are:

- To increase classification accuracy for fine-grained picture categories by refining and optimizing a Vision Transformer Model using the CIFAR-100 dataset.
- To examine the model's capacity to differentiate visually similar classes quantitatively using measures including accuracy, precision, recall, and confusion matrix analysis.
- To compare classification outcomes across related categories to examine how transformer-based design affects fine-grained categorization.

Next, the study is organized as follows. Section II is about related work on FGIC, and Section III describes our proposed system architecture. Section IV presents results and discussion. Finally, Section V discusses the conclusions of the work study and future work.

II. Related Work

This section will present an overview of the existing research that is related to Fine-Grained Image Classification and Transformer-based Vision models to determine the base of the proposed approach.

FGIC seeks to differentiate between visually similar subcategories, such as flower species or bird species, where distinguishing characteristics are frequently confined and subtle [7]. The initial approaches had limited flexibility and practicality since they mostly depended on hand-crafted features or part-based models with heavy supervision, such as bounding box and component annotations. Convolutional Neural Networks (CNNs) greatly enhanced feature representation with the development of deep learning; however, many CNN-based techniques continued to treat feature learning and discriminative domain identification as independent methods, leading to inadequate performance for fine-grained tasks.

This research proposed a system for enhancing self-supervised contrastive learning for fine-grained recognition problems called Fine-grained Adaptive Contrastive Learning (FACL) [8]. The method contributes to richer representation learning, which improves feature extraction, by including methods that provide increased sample variety. According to the experimental findings, FACL significantly outperforms conventional contrastive learning techniques. The approach demonstrated its ability to handle complicated visual differences with an accuracy of 64% in applications such as anomaly detection and fine-grained categorization. This emphasizes how important adaptive techniques are for contrastive learning, particularly when it comes to tasks that require identifying subtle differences.

Fine-grained categorization in agricultural disease detection, where categories only differ by minor images and are affected by background noise, humidity, and clarity [9]. The authors' Matrix-based Convolutional Neural Network (M-bCNN), which arranges convolutional layers in a parallel "kernel matrix", improves feature recognition. By expanding data streams, neurons, and link channels without significantly changing parameters, this setup outperforms traditional CNNs in feature representation. These methods enable the network to avoid vanishing gradients and overfitting. They produced a sizable, high-resolution dataset consisting of 16,652 photos of wheat leaves, which was expanded to 83,260 samples. One of the earliest large datasets for the classification of wheat leaf diseases is this one.

A hybrid architecture that combines the benefits of Vision Transformers (Vision Transformer Models) and Convolutional Neural Networks (CNNs) to improve FGIC [10]. The framework has three new modules. The Multi-Scale Image-to-Tokens (MIT) module extracts features at different scales to provide richer visual representations. The Mixed Convolution Feed-Forward (MCF) design enhances the network's ability to learn complex local properties. The Multi-Layer Feature Selection (MFS) method collects the most important attributes from many levels. This technique catches both broad and nuanced visual features by combining CNN's local spatial awareness with the Vision Transformer Model's global context knowledge. By enabling the model to recognize tiny differences between visually similar categories, this combination greatly increases classification accuracy in fine-grained identification tasks where tiny differences matter.

The current study addresses the problem of small inter-class changes in CIFAR-100, based on previous FGIC research that focuses on attention mechanisms and multi-scale feature learning. The proposed approach improves the model's capacity to differentiate closely related categories by utilizing Vision Transformers to capture both local and global images.

III. Materials and Methods

The proposed solution is an automated system designed to classify 100 various image classes efficiently and accurately. To achieve the classification tasks, the suggested system consists of a number of interconnected modules that operate one after the other. The CIFAR-100 dataset was utilized in this experiment. This dataset serves as the study's basis, as the suggested classification framework is trained and analyzed using its images.

Pre-processing is the initial module, which gets the input data appropriate for the model's architectural requirements. The

last module, the model module, receives the processed images and feeds them into the Vision Transformer architecture. The model extracts important characteristics, learns discriminative patterns across classes, and performs classification. This well-structured pipeline guarantees stable training, improved feature learning, and improved overall classification performance.

Fig. 2 describes the flow of the proposed solution. The first phase of the proposed solution is preprocessing. In order to stabilize the learning process, images were enlarged to the required resolution, normalized, and enhanced utilizing augmentation techniques such as random flipping and rotation. These techniques improve the model's ability to generalize, diversify the data, and reduce overfitting.

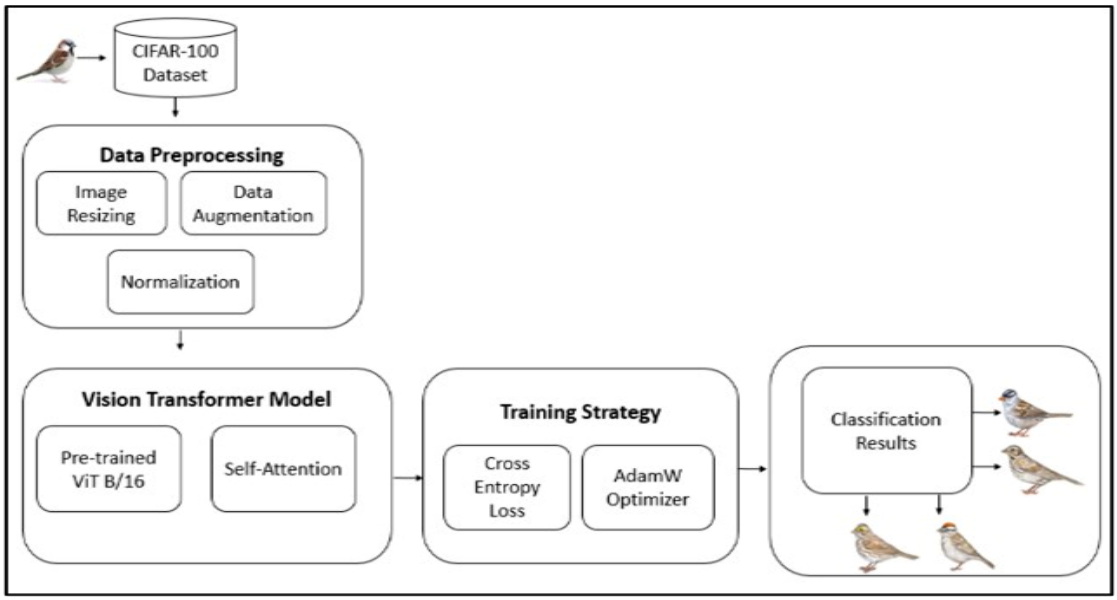


Fig. 2. An overview of the proposed methodology.

A. Image Resizing

Image resizing is one of the pre-processing stages that enhances the Vision Transformer Model by minimizing variations in image size. Vision Transformer Models with larger input sizes, which may divide pictures into meaningful patches, cannot represent the 32x32 pixel images in the CIFAR-100 dataset. This was fixed by up-sampling every image to 224 by 224 pixels. By scaling the pictures, the Vision Transformer Model may readily acquire patch embeddings that have both local and global characteristics. Learning fine details and being able to distinguish subtle visual differences across classes relies on these embedded images. The success of the classification. Fig. 3 shows the input and output images after resizing.

Eq. 1 presents a mathematical definition for this transformation, which involves sampling the original picture at scaled coordinates to determine the pixel intensity at spatial coordinates (x′, y′) in the stretched image. The scaled photos satisfy the Vision Transformer model's input size criteria while retaining structural consistency because of this method of normalizing the spatial dimensions.

Where H and W represent the image's height and breadth, and a pixel's intensity at spatial coordinates (x, y). All input photos are scaled to a fixed resolution of 224*224 in order to guarantee execution with the Vision Transformer Model architecture. The resizing process uses scaling based on the original height H and width W to transfer each pixel in the modified picture I′(x′,y′) to the appropriate place in the original image I(x',y).

$$I'(x', y') = I(x' H' . H, y' W' . W) \quad (1)$$

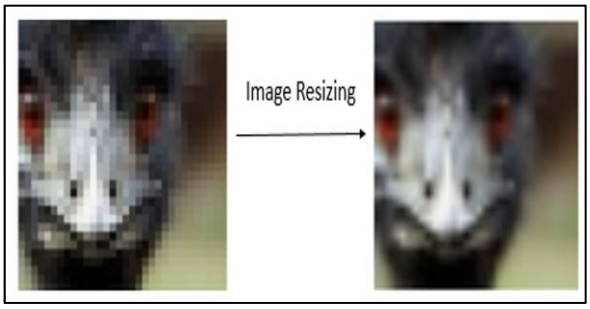


Fig. 3. Image resizing.

B. Data Augmentation

Data Augmentation in the proposed system was performed with torchvision. Data augmentation is a way to improve Vision Transformer frameworks for FGIC. The tiny images in CIFAR-100 data are barely changeable. We implement augmentation approaches to broaden the variety of the training samples. Random rotation, random cropping with padding, random horizontal flipping, and color jittering are common transformations. Image characteristics, including brightness, contrast, saturation, and hue, are affected by such changes. These modifications strengthened the model's resistance to changes in scale, orientation, and illumination.

The Vision Transformer Model was successfully trained utilizing a wider range of visual changes according to this approach. As a result, it improved overall categorization on fine-

grained categories and was better able to capture fine-grained visual distinctions. The data augmentation strategy, which enhances the model's performance on unknown test data and improves the classification rate. Fig. 4 clearly presents the Data Augmentation technique, which improves the performance of the model.

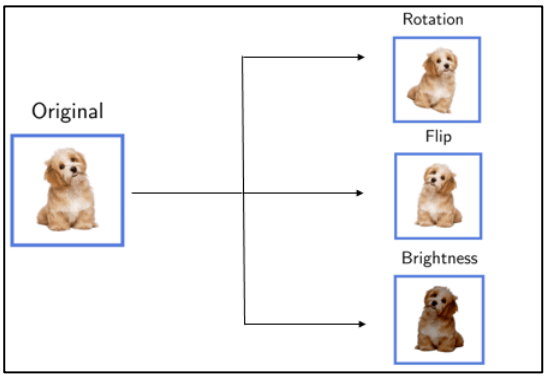


Fig. 4. Original and augmented images.

C. Random Horizontal Flip

A mirror image is produced across the vertical axis by this technique. This transformation is applied to training images at random, often with a 50% chance, in the context of Random Horizontal Flip.

In computer vision, this strategy is frequently used for data augmentation. It enhances a model's capacity for generalization by making it invariant to horizontal orientations, provided that it can accurately identify objects, regardless of whether they face left or right. To put it simply, every pixel along the horizontal axis is mirrored. The Random Flip Image is seen in Fig. 5. Each pixel at location (x', y) in the flipped image corresponds to the pixel at position (W−x−1, y) in the original image, according to Eq. 2.

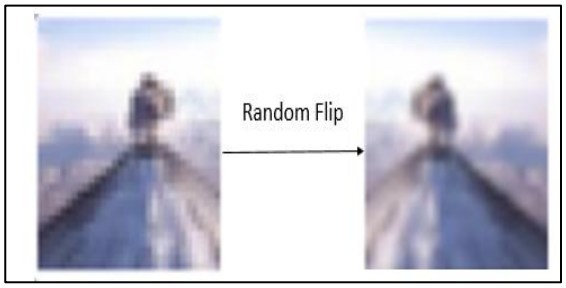


Fig. 5. Random flip image.

The expression I′ (x', y) =I (W−x−1, y) describes the mathematical operation for a horizontal flip of an image. Here, I (x', y) denotes the original image, I′ (x', y) is the transformed image after flipping, and W is the width of the image. Pixels from the left side move to the right, and pixels from the right side move to the left while the vertical position (y) stays the same.

$$I'(x,y) = I(W-x-1,y) \qquad (2)$$

D. Random Rotation

This method is frequently used to significantly enhance the range of training data in computer vision and machine learning. As a result, models become less dependent on variations in object placement and more stable. The algorithm learns to identify items despite their rotation in real-world scenarios by randomly rotating normal images during training. Eq. 3 describes a rotation transformation that is performed on the input image. The new coordinates (x', y') are obtained by rotating each pixel coordinate (x', y) by an angle θ. A typical 2D rotation matrix made up of sine and cosine terms is used to carry out the transformation. By introducing rotational invariance, this technique enhances the model's capacity to generalize to pictures with different angles. The Random Rotation is shown in Fig. 6.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (3)$$

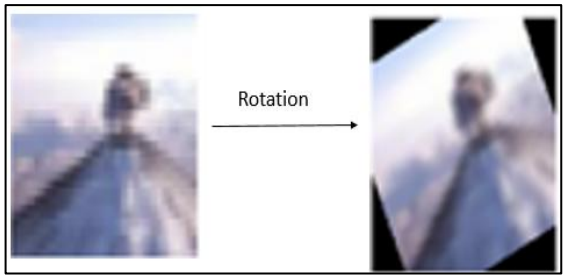


Fig. 6. Random rotation image.

E. Color Jitter Brightness Adjustment

This technique was used to enhance the robustness of image classification models by simulating various lighting conditions. Fig. 7 illustrates the Color Jittering Brightness Adjustment. Eq. 4 describes a brightness adjustment applied to the input picture, where a random brightness offset Δ (b) is added to the original image I to generate the transformed image *I*. To generate a new image, the input image's brightness is randomly changed during this procedure. This fluctuation reflects real-world situations where brightness may vary depending on the setting or time of day. The model's capacity to generalize to new data is enhanced by training on standard images with varying brightness levels, which reduces the model's sensitivity to lighting variations.

$$I' = I + \Delta_{b,} \quad \Delta_b \sim \cup(-\beta, \beta) \qquad (4)$$

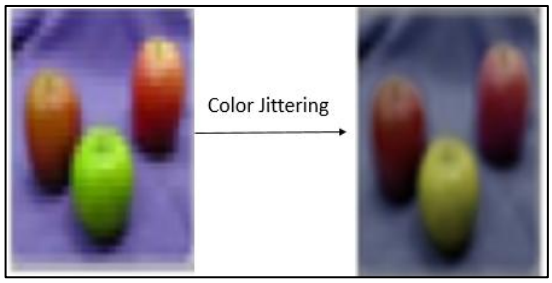


Fig. 7. Input and enhanced image with color jittering.

F. Contrast Adjustment

The operation changes the brightness difference between light and dark areas of an image. This makes features easier or harder to see. Eq. 5 describes a contrast adjustment operation applied to the input image, where the pixel intensities of the transformed image I' are scaled by a factor α around the mean intensity μ. This random contrast adjustment is often used as a data augmentation method in image classification tasks. It helps a model handle different lighting and image quality conditions better. Fig. 8 shows color illumination correction of the input image.

$$I' = \alpha \cdot I - \mu + \mu, \quad \alpha \sim \cup\ (1 - c\ , 1 + c) \quad (5)$$

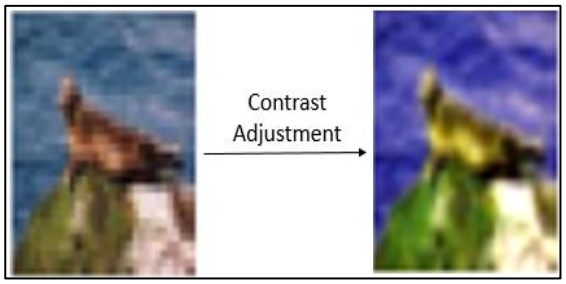


Fig. 8. Color illumination correction of the input image.

G. Normalization

This is the second pre-processing step. Images were resized, and pixel values were scaled so that each color channel's mean value and standard deviation were 0.5. Eq. 6 normalizes the input values to a fixed range, usually between -1 and 1. It helps to speed up and stabilize the training process. Normalization ensures that the pre-trained Vision Transformer model predicts the input distribution. Better convergence and performance assurance result from this. Fig. 9, which displays the Enhanced Image after Normalization, illustrates the difference gained after applying the augmentation. By eliminating internal variance shift, normalization contributes to more effective and efficient learning during training. This process is typically subject to the normalization formula of the standard score (z- score) of each channel of an image (Red, Green, Blue) [Eq. (6)].

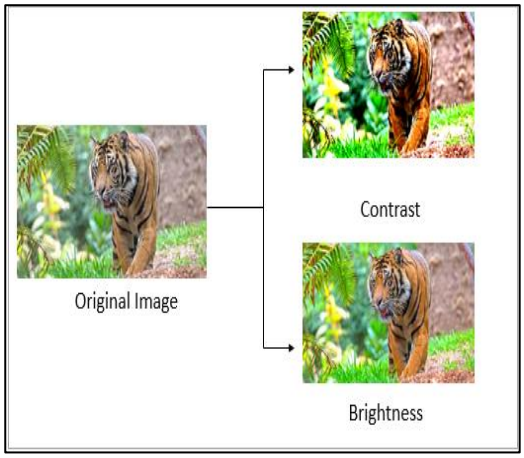


Fig. 9. Original and enhanced images.

$$X' = \frac{X - \mu}{\sigma} \quad (6)$$

where, X represents the original pixel value, μ is the mean, and σ is the standard deviation. The mean and standard deviation values used are both 0.5 for each color channel.

H. Fine-Grained Image Classification

The Fine-Grained Image Classification Model, utilizing the Vision Transformer Model, is employed in the proposed method. The strategy is based on the Vision Transformer architecture, which has shown outstanding outcomes in various types of image classification tasks. Unlike traditional convolutional neural networks, the Vision Transformer model employs self-attention processes to identify global characteristics and long-range interactions within the input image. The main architecture used in this study is the Vision Transformer Model-Base-Patch16-224 timm (PyTorch Image Models) library. This approach divides the image provided into many non-overlapping patches. These patches are then passed through a typical transformer encoder after being encoded as fixed-length vectors. Each of the 12 transformer layers in the Vision Transformer Model-Base-Patch16-224 contains 12 attention heads and 768 hidden dimensions.

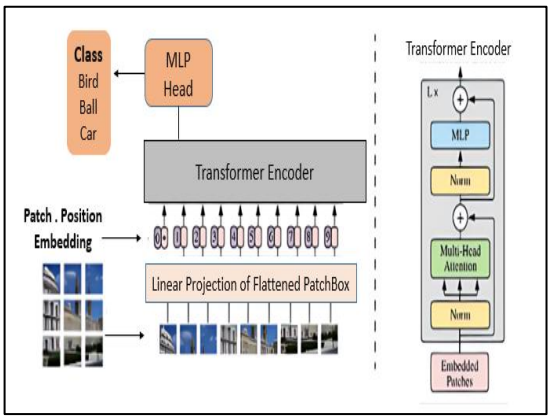


Fig. 10. Vision transformer architecture [11].

Fig. 10 illustrates that the Vision Transformer Model works with images by splitting them into patches of a fixed size. The patches are flattened and transformed into a numerical embedding representing the visual content of the patch. These embeddings are then sent into the transformer encoder together with positional encodings and a learnable class token. The encoder uses feed-forward layers with residual and normalization connections as well as multi-head self-attention. This structure helps the model recognize spatial data in images by generating significant global relationships between the patches.

A Multi-Layer Perceptron (MLP) head receives the class token's final output and is used to perform the classification task. The Vision Transformer Model can effectively capture long-range connections with this framework. The model is effective, particularly when handling recognition of images and fine-grained classification tasks, since the transformer attention mechanism allows the model to learn to focus on tiny images. Recognizing related groups of objects is essential, particularly when a collection includes sensitive classifications. The backbone may be used to save training time and enhance the

model's generalization. It makes it possible to determine tiny variations across classes in fine-grained visual datasets.

The proposed solution for FGIC:

1) The Vision Transformer (Vision Transformer Model) model uses ImageNet-pretrained weights to improve performance through transfer learning, and it was modified for the CIFAR-100 classification problem. Faster convergence and increased accuracy are made possible by this initialization, which offers a solid foundation of learnt visual features, especially in fine-grained classification settings with less training data. The technique successfully addresses the problem of data scarcity by transferring rich visual information to the target domain. Furthermore, the framework uses PyTorch to provide effective computing in both CPU and GPU settings, dynamically assigning model parameters to the available hardware.

2) The pre-processed and augmented CIFAR-100 dataset is used for training after the Vision Transformer (Vision Transformer Model) has been initialized and configured. Using a cross-entropy loss function appropriate for multi-class tasks, model parameters are optimized to minimize classification errors across 100 object categories. For transformer-based architectures, the optimization process uses the AdamW optimizer, which offers better weight regularization than standard Adam. A learning rate of 3e-5 guarantees gradual convergence, while a weight decay of 0.01 aids in reducing overfitting and enhancing generalization. Furthermore, the learning rate is gradually reduced using a cosine annealing learning rate scheduler, which facilitates smoother convergence and lowers the possibility of becoming stranded in local minima.

3) Monitor loss and accuracy metrics during training. This helps the model develop useful visual characteristics. To ensure effective learning, hyperparameters were adjusted as required. Once the training is completed, the Vision Transformer will be capable of generating good feature representations that allow the system to perform correct classification in the fine-grained disjointed categories in the CIFAR-100 dataset.

4) After training, the Vision Transformer Model's ability to generalize is determined using the test split of the CIFAR-100 dataset, with classification accuracy functioning as the primary measure. A confusion matrix is examined to provide a deeper understanding of performance. This analysis highlights misclassifications and reveals class-wise prediction patterns, particularly among visually similar categories with overlapping attributes like texture, color, and shape. This analysis helps in highlighting the model's problems in fine-grained parameters and offers recommendations for additional optimization, such as enhancements to data augmentation, class balancing, and architectural modifications, which will ultimately improve the model's capacity to differentiate closely related classes and boost overall performance.

IV. Results and Evaluation

The proposed Vision Transformer-based method for FGIC is evaluated on the CIFAR-100 dataset, which shows how well it learns discriminative visual characteristics and achieves consistent generalization performance. To achieve a thorough evaluation, the model was evaluated using a variety of performance indicators, including test accuracy, precision, recall, F1-score, and confusion matrix analysis. Accuracy, macro precision, macro recall, and macro F1-score were used to assess the recommended Vision Transformer Model's performance on the CIFAR-100 dataset. The model demonstrated good prediction abilities across 100 item categories, with an overall classification accuracy of 89.68% on the test set. The findings utilizing Vision Transformer Model B/16 on CIFAR-100 are shown in Fig. 11.

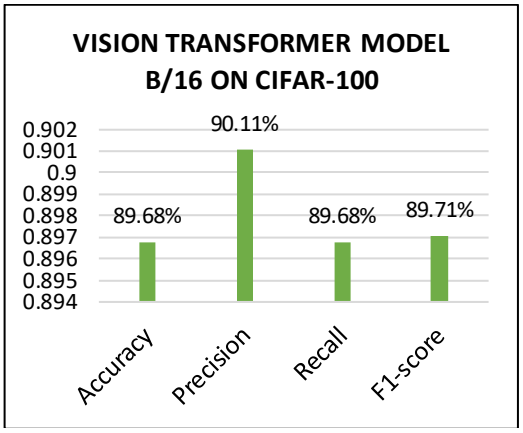


Fig. 11. Evaluation of proposed solution using vision transformer model B/16 on CIFAR-100.

By evaluating both class-wise performance and overall prediction accuracy, these metrics offer a comprehensive picture of the model's performance. Combining accuracy with precision, recall, and F1-score provides a broader overview of the model's performance in recognizing certain categories over the whole dataset.

A. Analysis of Test Accuracy

The Vision Transformer Model's generalization performance on the CIFAR-100 dataset can be enhanced by analyzing test accuracy over epochs. The model shows a sharp increase in test accuracy in the early epochs. This sharp rise indicates that the pretrained Vision Transformer Model backbone quickly adapts to the CIFAR-100 classification task and learns meaningful feature representations within a few training iterations. In transfer learning circumstances, where pretrained weights speed up learning, early convergence is frequently seen. Overall, the test accuracy trend shows that the model reaches its best generalization performance in the early epochs and continues to provide reliable outcomes beyond that. The model test accuracy results are displayed in Fig. 12. The model achieves about 89.68% test accuracy on CIFAR-100, demonstrating continuous generalization ability across epochs.

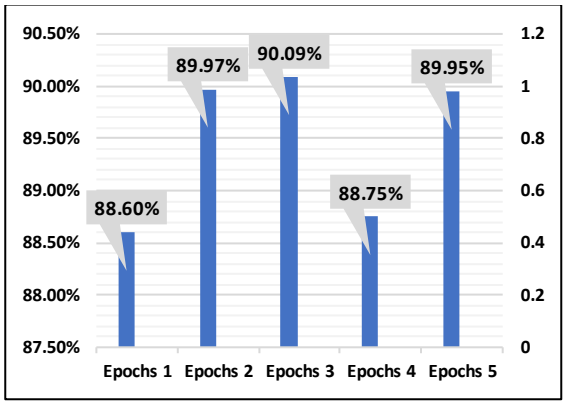


Fig. 12. Model test accuracy results.

B. CIFAR-100 Per-Class Precision Results

The proposed Vision Transformer Model's per-class accuracy in all 100 CIFAR-100 dataset classes. Most classes have high accuracy values, which often fall between 0.85 and 1.00, meaning that the model is typically accurate when predicting a certain class. This indicates that the model has a high capacity to reduce false positive predictions across various item types. For the majority of classes, consistent predictive reliability is shown by the very small range in bar heights. However, other classes show significantly better accuracy, with values lying closer to 0.70–0.80.

This consistency supports the effectiveness of the Vision Transformer architecture in handling multi- class image recognition on CIFAR-100. Fig. 13 shows that the per-class results typically show variability across different categories, indicating that some classes achieve higher precision.

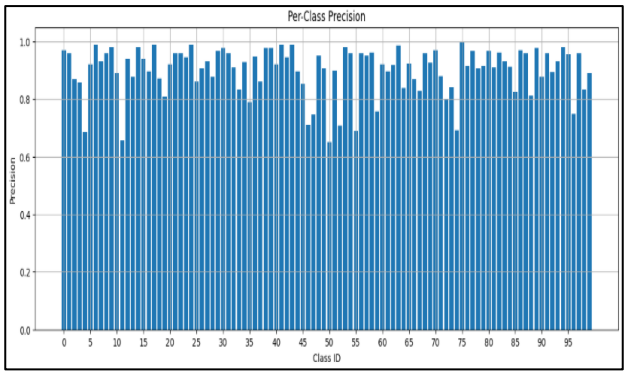


Fig. 13. Per-class precision results.

C. CIFAR-100 Per-Class Recall Results

The proposed Vision Transformer model's recall performance for each of the 100 CIFAR-100 classes is shown in the per-class recall graph. The majority of classes have high recall values, usually between 0.85 and 1.00, which suggests that the model correctly identifies a significant portion of real examples that correspond to each class. This shows how well the model can reduce false negatives and accurately gather relevant data across a variety of object classifications. Stable detection performance across the dataset is shown in the bars' overall consistency. Recall levels for a few classes are significantly lower, falling closer to around 0.70–0.80. The per-class recall is shown in Fig. 14.

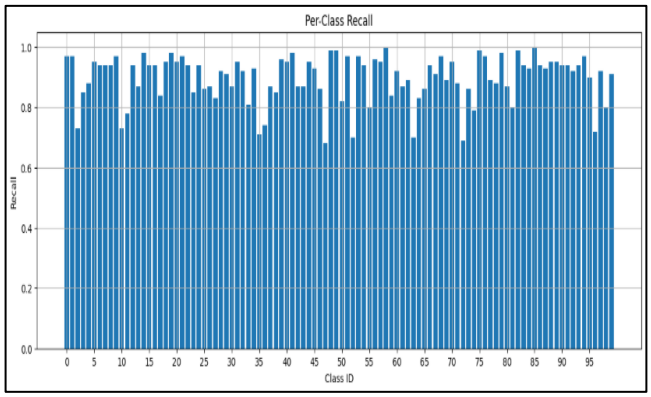


Fig. 14. Per-class recall results.

Overall, the distribution demonstrates how effectively the Vision Transformer-based architecture handles challenging multi-class image classification tasks on the CIFAR-100 dataset.

D. Confusion Matrix Analysis

The confusion matrix displays a strong diagonal pattern, indicating that the majority of samples are properly identified, providing a deeper understanding of class-wise performance beyond total accuracy. This illustrates how well the model captures significant visual patterns and how resilient the taught features are.

Despite this, the model performs well across the majority of classes, and the small number of mistakes indicates that misclassifications are not common. All things considered, the attention-based Vision Transformer (Vision Transformer Model) successfully captures both local and global features, producing dependable and balanced classification performance with significant room for improvement through improved feature refinement. Fig. 15 provides a class-by-class evaluation of the proposed model's prediction performance.

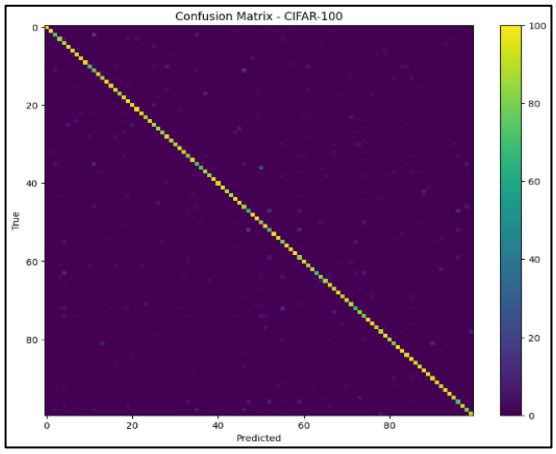


Fig. 15. Confusion matrix of predicted classes.

A major problem in fine-grained picture classification, because of minute variations in form, texture, and color, is that, although some misclassifications occur in off-diagonal regions, they are primarily limited to classes with similar visual features.

In datasets like CIFAR-100, where low resolution and closely related categories make classification challenging, this type of confusion is especially visible.

V. Comparative Analysis with Existing Research

Using the CIFAR-100 dataset, three sample image classification and few-shot learning techniques are compared in this section. Model architecture, learning methodology, and applicability for the CIFAR-100 dataset FGIC difficulties are the main areas of comparison. On the CIFAR-100 dataset, existing methods have shown excellent performance on challenging picture classification tasks. In order to increase recognition when training data is limited, several techniques concentrate on improving feature representation by utilizing attention processes and multi-branch structures to capture subtle correlations among visual characteristics. Other strategies focus on enhancing model generalization by using structured regularization techniques that hide spatially correlated areas in feature maps to avoid overfitting and enhance classification performance on multi-class datasets. Fig. 16 shows the graphical representation of the comparative analysis of the proposed solution with existing research on the CIFAR-100 Dataset.

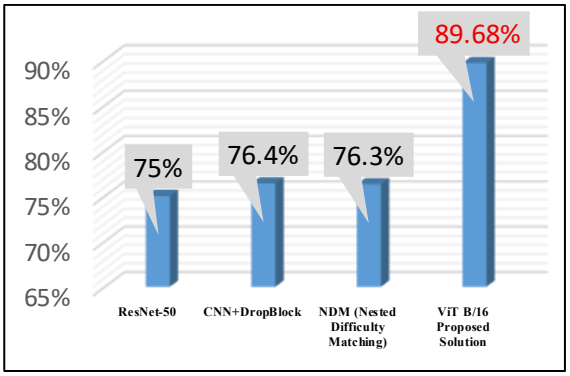


Fig. 16. Comparative analysis with existing research on CIFAR-100.

Additionally, dataset distillation techniques have been proposed to generate compact yet informative synthetic datasets by prioritizing difficult training samples and aligning them with different stages of the learning process, enabling models to achieve performance close to full-dataset training. However, many of these approaches rely on convolutional neural network backbones and require additional modules to enhance feature learning. In comparison, transformer-based architectures such as the Vision Transformer Model naturally capture global relationships between image patches through self-attention mechanisms, making them particularly suitable for FGIC, where accurate recognition depends on identifying subtle differences between visually similar categories.

VI. Conclusion and Future Work

Using the CIFAR-100 dataset, this study effectively developed and examined a Vision Transformer-based framework for FGIC. The proposed framework performed well in macro-level precision, recall, and F1-score, yielding stable and reliable results across all 100 categories, with an accuracy of 89.68%. The overall results highlight the potential of the Vision Transformer Model for FGIC despite obstacles like the dataset's low resolution and high computing needs. The findings imply that transformer topologies can be a reliable and effective method for applications requiring complex and subtle image distinction.

Future studies will expand the assessment to benchmark fine-grained datasets, including FGVC-Aircraft, Stanford Cars, and CUB-200-2011, in order to validate the suggested model more thoroughly in fine-grained image classification, even though the Vision Transformer showed strong performance on the CIFAR-100 dataset. In order to take advantage of both local and global feature representations, future research may potentially look at hybrid architectures that mix convolutional neural networks with vision transformers. Additionally, using few-shot or self-supervised learning techniques might lessen the need for sizable labeled datasets and enhance the model's capacity for generalization in practical situations.

Acknowledgment

The authors extend their appreciation to the Arab Open University for funding this work.

References

[1] Y. Liu *et al.*, "Deep learning for fine-grained image analysis: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] W. Ge, X. Lin, and Y. Yu, "Weakly supervised fine-grained classification with part selection from CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014. doi: 10.5244/C.28.1.

[4] J. Krause, M. Stark, J. Deng, and Li Fei-Fei, "3D object representations for fine-grained categorization," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), 2013.

[5] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," arXiv preprint arXiv:1306.5151, 2013.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Institute of Technology, 2011.

[7] Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] T. Yin, J. Wang, Y. Zhao, H. Wang, Y. Ma, and M. Liu, "Fine-grained adaptive contrastive learning for unsupervised feature extraction," *Neurocomputing*, vol. 618, p. 129014, 2025, doi: 10.1016/j.neucom.2024.129014.

[9] Lin, Q. Chen, Y. Huang, and X. Wang, "Matrix-based convolutional neural network for fine-grained wheat disease recognition," *IEEE Access*, vol. 7, pp. 123450–123462, 2019.

[10] R. Shao, X.-J. Bi, and Z. Chen, "Hybrid Vision Transformer Model-CNN network for Fine-Grained Image Classification," *IEEE Signal Processing Letters*, vol. 31, pp. 1109–1113, 2024, doi: 10.1109/LSP.2024.3386112.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint* arXiv:2010.11929, 2021.

MRAE: Multi-Resolution Attention Ensemble with Hybrid CNN–Transformer Fusion for Breast Ultrasound Classification

Hemin Kareem Azeez Alshateri, Ahmed Harbaoui
Department of Computer Science, King Abdulaziz University, Jeddah, KSA

***Abstract*—Breast ultrasound images can be classified as benign, malignant, and normal. Due to the imbalanced distribution of classes in breast ultrasound images, intra-class heterogeneity of lesions, and ultrasound artifacts like speckle noise, classification of breast ultrasound images remains a challenging problem. In this paper, we present MRAE, a hybrid architecture with a DenseNet121 convolutional encoder and two transformer encoders (ViT-Base and DeiT-Base-Distilled). These branches are run in parallel with input sizes of 192×192 pixels, 224×224 pixels, and 256×256 pixels, respectively. The learned feature representations from these branches are fused using a cross-attention block and combined using learnable ensemble weights. Focal loss with deep supervision is used during training along with CutMix regularization, Weighted Random Sampling, and Cosine Annealing. We perform experiments using 10-fold stratified cross validation on the benchmark BUSI dataset (780 Images). MRAE achieves an average accuracy, macro F1-score, and macro recall of 93.72%, 94.25%, and 95.02%, respectively, across all cross-validation folds. The ResNet50 baseline achieves accuracy, F1-score, and recall of 90.64%, 91.36%, and 91.62% across all folds. We show that MRAE has significantly lower standard deviations across cross folds, indicating better stability. Our method provides evidence that breast ultrasound images can be classified accurately and reliably in a multi-resolution attention fusion network for use in clinical breast cancer screening.**

***Keywords*—Breast ultrasound classification; multi-resolution learning; CNN–Transformer fusion; cross-attention ensemble; Vision Transformer (ViT)**

I. Introduction

Breast cancer is one of the most common and deadly cancers among women. An early and accurate diagnosis of breast cancer allows physicians to treat the patient timely which can increase the survival rate [1]. Breast ultrasound is one of the most commonly used techniques for diagnosis because it is readily available, inexpensive [2], and does not use ionizing radiation [3]. Manual interpretation of breast ultrasound images is difficult due to their subjective nature and could vary from observer to observer [4]. Automated classification of breast ultrasound images using deep learning can be used as assistance for the clinicians to avoid misclassification [5].

CNNs have achieved promising results in medical image analysis by effectively encoding local texture and spatial context information [6]. Recently, ViTs and their derivatives have exhibited a complementary capability of encoding information using self-attention to extract long-range global dependencies [7]. However, previous methods are mostly single-model frameworks that use a fixed-resolution input. Due to this limitation [8], the learned representations often lack richness and diversity. In addition, class imbalance is an issue not fully addressed in prior works [9]. Since malignant and normal findings are underrepresented compared to benign findings in most clinical datasets, class-imbalanced learning can result in classifiers that are biased towards the majority class [10].

Most methods cannot fully leverage multi-scale feature representations and also fail to sufficiently incorporate a cross-architecture feature fusion scheme that emphasizes more robust features [11]. In addition, they do not address class imbalance issues presented in small-to-medium sized clinical datasets. Furthermore, consistency of model performance across heterogeneously partitioned data is not well explored [12].

To fill these shortcomings, we introduce MRAE (Multi-Resolution Attention Ensemble), a mixed CNN-Transformer model that ensembles DenseNet121, ViT-Base, and DeiT-Base learning at different resolutions through cross-attention layers and learnable ensemble weights. Complemented with focal loss with auxiliary deep supervision, CutMix augmentation, and weighted sampling to alleviate class imbalance, MRAE attains new state-of-the-art performance in terms of accuracy (93.72%), F1-score (94.25%), and recall (95.02%) with significantly enhanced cross-fold consistency over the ResNet50 baseline.

II. Related Work

Efforts have been made recently to classify breast ultrasound images automatically. Methods such as fine-tuning pre-trained CNN models and ensembles of CNN and Transformer models have been explored. Kormpos et al. [13] reviewed several deep learning models for breast tumor classification with transfer learning on BUSI dataset. The authors used a cascaded classification pipeline and found that vision Transformer models that use an attention mechanism can be used as an alternative to popular CNN models. Limitations, including model explainability and high computational resource requirements, remain. Architectures like InceptionV3 provide competitive baselines but are limited by being single models operating at a fixed input resolution.

To overcome the problem of local and global features complementarity, Asif et al. [14] designed a feature fusion-based deep learning model consisting of MobileNetV2 and DenseNet121 coupled with attention mechanisms to perform benign-malignant classification on a private clinical dataset of 2171 images, as well as the public BUSI dataset, where they

were able to reach an AUC of 0.9834 on the public benchmark. Results showed that fusion of advanced features coupled with attention mechanisms drastically outperformed CNN single-architecture baselines and surpassed the diagnostic performance of radiologists with varied levels of experience. However, this DL model is a binary classifier and does not account for the clinically relevant three-class setting that includes normal tissue.

In similar attempts to learn more explainable multi-scale representations, Saini et al. [15] proposed a variational mode decomposition-guided CNN architecture for breast lesion classification in ultrasound images, using 2D-VMD [34] as a key component to learn self-explanatory lesion-specific boundary and texture maps to drive a mixed pooling and attention-guided CNN model for breast ultrasound lesion classification with comparable accuracy and intrinsic explainability of learned features on the BUSI dataset [15]. However, this method does not make use of a transformer backbone, which may have higher global receptive fields for modeling long-range context. On the other hand, Yıldırım et al. [16] recently proposed a ViT-based ensemble learning framework for three-class breast ultrasound classification using the BUSI dataset. ROI segmentation is performed on the training set images along with Albumentations-based image augmentation before being fed into multiple transformer backbones that are ensembled by simple averaging to boost performance. The authors show that ensembling yields better classification performance with higher accuracy, precision, and recall than individual transformers. However, this method still does not leverage multi-resolution input or cross-attention CNN-transformer fusion schemes and does not explicitly deal with class imbalance via loss-weighting.

III. Proposed Methodology

In this section, we will introduce all the details of our MRAE framework, including the dataset, preprocessing pipeline, our proposed model and training strategy. Each step is carefully designed to tackle one of the aforementioned challenges, including class imbalance problem, lack of multi-resolution learning and lack of cross-architecture feature fusion for breast ultrasound classification. The methodology has been visualized in Fig. 1.

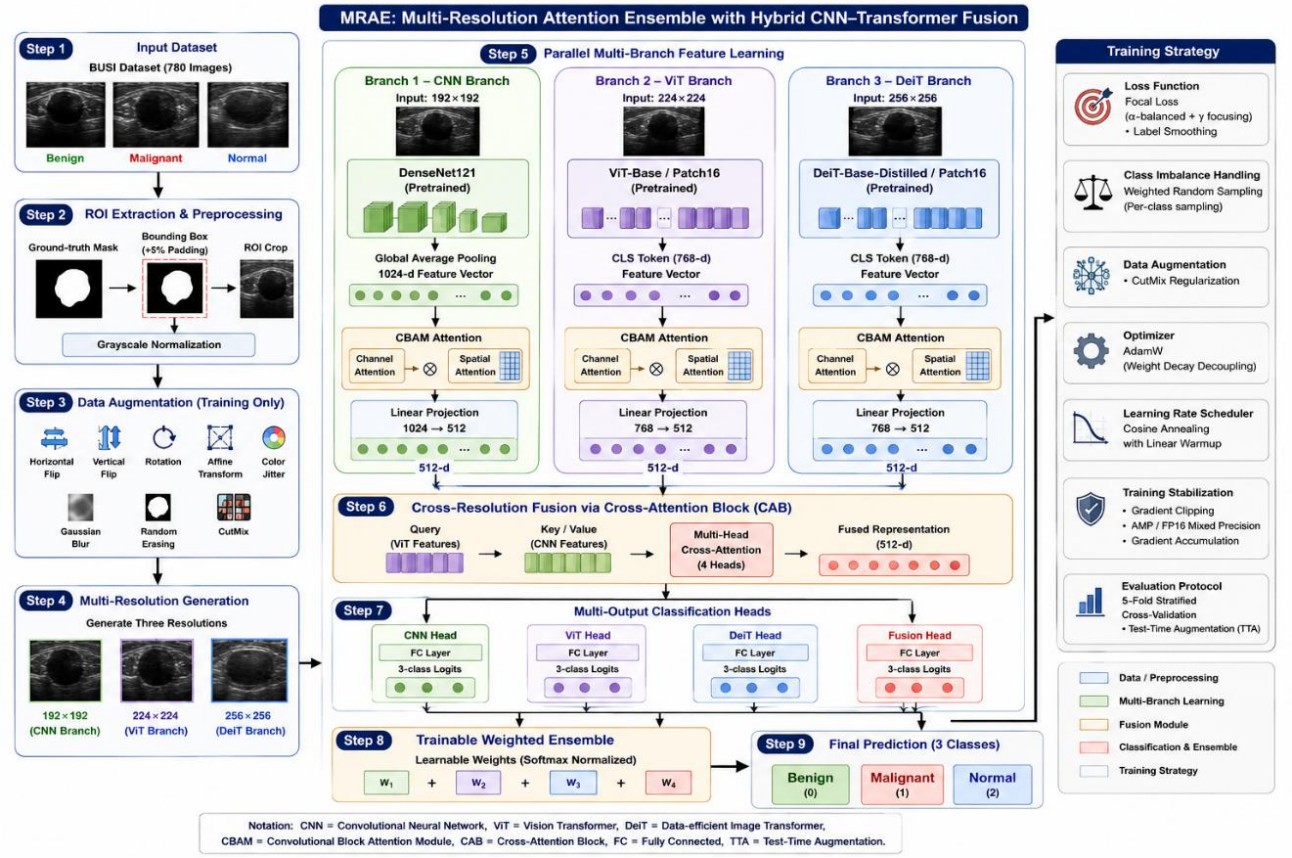


Fig. 1. Flowchart for the proposed methodology.

A. Dataset

We use the Breast Ultrasound Images (BUSI) dataset which is a commonly used public benchmark for breast lesion classification and segmentation tasks [17]. The dataset consists of 780 grayscale ultrasound scans of breasts acquired from 600 female patients. The images are categorized into three classes: benign (437 images, 56.0%), malignant (210 images, 26.9%) [18] and normal (133 images, 17.1%). Each scan has one or more pixel-wise segmentation masks provided which mark the region occupied by lesion/s. The dataset, therefore contains precise ground-truth segmentation masks suitable for supervised training. Images with more than one mask represent overlapping lesions acquired from the same scan. In 18 benign cases, there

are multiple masks. In addition, the width and height of images also vary due to natural variation of ultrasound scans. All images are resized to the longest side of 916 pixels and saved as RGB images. Width of the images varies from 323 pixels to 916 pixels. Height of the images varies from 393 pixels to 716 pixels. The BUSI dataset has a high-class imbalance with benign category having over half the data. There are three times as many benign cases than normal cases (the least prevalent class). Class imbalance is problematic since most classification models are biased towards the majority classes. In medical applications like this where the negative classes (malignant and normal) are very important, class imbalance can lead to low sensitivity for these categories. Therefore, overcoming class imbalance was one of the design considerations while developing our proposed framework. We mitigate this problem by using weighted sampling during training. We also employ focal loss as our loss function as it performs well on unbalanced datasets, as shown in Table I [19].

TABLE I. BUSI DATASET SUMMARY

Category	Number of Images	Percentage	Annotation Type
Benign	437	56.0%	Pixel-level segmentation mask
Malignant	210	26.9%	Pixel-level segmentation mask
Normal	133	17.1%	No mask (no lesion)
Total	780	100%	—
Patients	600	—	—
Female patients only	—	—	—
Image Width Range	323–916 px	—	Grayscale (stored as RGB)
Image Height Range	393–716 px	—	Grayscale (stored as RGB)
Multi-mask Cases	18 (benign)	2.3%	Multiple lesions per scan

B. Preprocessing and Data Preparation

Input images are preprocessed using a multi-step pipeline before standardization and feeding them into the model. The aim is to present the model with data that ensures comparability and clinical meaningfulness of features. Benign and malignant scans undergo region-of-interest (ROI) extraction based on their segmentation mask [20]. The bounding box is calculated based on each lesion's mask and expanded by 5% to include surrounding tissue. This allowed us to decrease background noise while maintaining clinically relevant perilesional information. Normal scans are not annotated with any lesion and therefore are not cropped. In cases where multiple masks were present for one image, all masks were combined using element-wise maximum to generate a lesion mask that contains all lesions. Then, images undergo resizing to three different resolutions at once: 192 × 192 pixels for the CNN branch, 224 × 224 pixels for the ViT branch and 256 × 256 pixels for the DeiT branch [21,22]. This allows each branch of the network to extract slightly different features based on varied resolutions. Training augmentation is then applied individually per resolution with random horizontal and vertical flip (probability, $p = 0.5$), random rotation ($\pm 20°$, $p = 0.5$), and finally random application of Gaussian noise or Gaussian blur ($p = 0.3$). Images are then normalized (mean = 0.5, std = 0.5) for each channel. At validation and inference time, no augmentations were used besides resizing and normalization. We used Weighted Random Sampler for training to ensure that classes are proportionally sampled for every mini-batch, as shown in Table II.

TABLE II. PREPROCESSING PIPELINE SUMMARY

Step	Operation	Parameters	Applied To
ROI Extraction	Bounding box from segmentation mask	5% buffer on each side	Benign & Malignant only
Multi-mask Merging	Element-wise maximum of all masks	—	Multi-lesion cases only
Resize — CNN Branch	Bilinear resizing	192 × 192 px	All images
Resize — ViT Branch	Bilinear resizing	224 × 224 px	All images
Resize — DeiT Branch	Bilinear resizing	256 × 256 px	All images
Horizontal Flip	Random flip	p = 0.5	Training only
Vertical Flip	Random flip	p = 0.5	Training only
Random Rotation	Rotation	±20° (p = 0.5)	Training only
Noise / Blur	Gaussian noise or blur	p = 0.3	Training only
Normalization	Mean & Std normalization	Mean = 0.5, Std = 0.5	Training & Validation
Weighted Sampling	Inverse-frequency class weights	Per-class	Training only

C. MRAE Model Architecture

TABLE III. MRAE ARCHITECTURE COMPONENTS

Component	Backbone / Module	Input Resolution	Output Dimension	Pretrained
CNN Branch	DenseNet121	192 × 192 px	1024-d	ImageNet
ViT Branch	ViT-Base / Patch16	224 × 224 px	768-d	ImageNet
DeiT Branch	DeiT-Base-Distilled / Patch16	256 × 256 px	768-d	ImageNet
CNN Projection	Linear layer	1024-d	512-d	—
ViT Projection	Linear layer	768-d	512-d	—
DeiT Projection	Linear layer	768-d	512-d	—
Cross-Attention Block	Multi-head attention (4 heads)	ViT (Q), CNN (K, V)	512-d	—
Residual + LayerNorm	Post-attention normalization	512-d	512-d	—
Classification Heads	Linear (×3, one per branch)	512-d	num_classes = 3	—
Ensemble Fusion	Learnable softmax weights (w1, w2, w3)	3 × logits	Final prediction	—
Baseline (ResNet50)	ResNet50 + Dropout(0.3) + Linear	224 × 224 px	num_classes = 3	ImageNet

Table II depicts an overview of our suggested MRAE architecture, which consists of three branches with varying input

resolutions. These branches are complementary and are fused together using cross attention and learnable ensemble weights. The CNN branch consists of a pretrained DenseNet121 model used as a convolutional backbone with the classifier head removed and replaced with an identity layer. Features from the final dense block are taken after adaptive average pooling is applied, forming a 1024-dim embedding from a 192×192 input. The vision transformer branch utilizes a pretrained ViT-Base/Patch16 model. This gives a 768-dim class embedding of a 224×224 sized input. Lastly, our third branch features a pretrained DeiT-Base-Distilled/Patch16 transformer with flexible image sizes enabled. This forms our 768-dim feature vector from a 256×256 input. Since each ViT-B/16 backbone contains approximately 86 million trainable parameters, the complete MRAE model contains approximately 260 million trainable parameters in total, making it substantially larger than conventional CNN-based architectures while enabling richer multi-resolution feature learning.

Embedding space using three linear projection layers. The Cross Attention Block (CAB) then attends to the convolutional features from the CNN branch using the projected ViT features as a query, with the CNN projected features used as a key and value. This operation consists of a four-head multi-head attention module followed by a residual connection and Layer Normalization. This allows the transformer model to attend to spatially relevant features from the convolutional branch. The three branches each have their own classification head that provides individual class logits. The ensemble prediction is then made by taking a softmax over the weighted sum of the three branches. These weights are learnable and initialized to be equal, and trained end-to-end with the rest of the network. The 192×192, 224×224, and 256×256 inputs are chosen so that the image representations learned at each scale provide a complementary, multi-scale view of the US imagery. The features learned at lower resolutions force the model to represent the information contained in small regions of the image (local textures, speckle distribution, edges of lesions). In contrast, representations learned at higher resolutions help the model retain contextual information about the larger-scale structures and the spatial relationships between them. Features learned at 224×224 resolution fall naturally in between those learned at 192×192 and 256×256 resolution. DenseNet121 provides the best ability to extract local texture features with limited data by taking advantage of its highly redundant feature reuse. ViT-Base is able to learn global image dependencies since self-attention is able to model global dependencies (something CNNs are not able to do). Finally, DeiT-Base-Distilled allows for more data-efficient learning and a more stable training process. Its inherent knowledge distillation strategy allows it to better generalize to smaller datasets like BUSI.

D. Training Procedure and Evaluation Protocol

The MRAE model was optimized with a cross-entropy loss, also known as Focal Loss, with class-balancing weights [1.0, 2.0, 1.0] for the benign, malignant, and normal classes, respectively, and focal loss hyperparameter gamma of 2.0. Classification loss contributions from each individual branch, known as auxiliary loss terms, were added to the ensemble objective. Total loss during training was a combination of the ensemble focal loss plus 0.3 times each branch-level loss. These auxiliary losses apply deep supervision to ensure each branch learns independent discriminative features. Training employed CutMix augmentation with alpha=1.0 and probability p=0.5. Mixed pairs of training samples and their labels were linearly interpolated with a mixing coefficient sampled from Beta(1,1) and applied identically to all 3 resolutions. AdamW optimization was performed with a learning rate of 2×10^{-4}, weight decay of 1×10^{-4}, cosine learning rate scheduling across 50 epochs. Gradient accumulation over 4 steps was used to reach an effective batch size of 16 (with a physical batch size of 4), as shown in Table IV.

TABLE IV. TRAINING CONFIGURATION AND EVALUATION PROTOCOL

Setting	MRAE	Baseline (ResNet50)
Optimizer	AdamW	AdamW
Learning Rate	2×10^{-4}	1×10^{-4}
Weight Decay	1×10^{-4}	1×10^{-4}
LR Scheduler	Cosine Annealing (T_max = 50)	Cosine Annealing (T_max = 15)
Max Epochs	50	15
Early Stopping Patience	8 epochs	4 epochs
Physical Batch Size	4	4
Effective Batch Size	16 (4 accumulation steps)	16 (4 accumulation steps)
Mixed Precision	AMP / FP16	AMP / FP16
Loss Function	Focal Loss (γ = 2.0)	Focal Loss (γ = 2.0)
Class Weights (α)	[1.0, 2.0, 1.0]	[1.0, 2.0, 1.0]
Auxiliary Branch Loss	0.3 × (CNN + ViT + DeiT losses)	Not applicable
CutMix Augmentation	p = 0.5, Beta(1,1)	Not applied
Test-Time Augmentation	Original + horizontal flip (averaged)	Original + horizontal flip (averaged)
Evaluation Protocol	10-fold Stratified Cross-Validation	10-fold Stratified Cross-Validation
Primary Metrics	Accuracy, Macro F1, Macro Recall	Accuracy, Macro F1, Macro Recall
Best Model Selection	Highest validation F1 per fold	Highest validation F1 per fold

Automatic Mixed Precision training was used. Early stopping with patience of 8 epochs was used to restore the checkpoint with the best validation F1 score. Test-time augmentation averaged the predictions between the original image and the horizontal flip. Training was performed with 10-fold Stratified Cross-Validation. Accuracy, macro-averaged F1-score, and macro-averaged recall are reported as mean ± standard deviation across folds.

IV. RESULTS AND DISCUSSION

We report the quantitative experimental results of our proposed MRAE framework against the baseline ResNet50 using 10-fold stratified cross validation in this section. Results are discussed in terms of Accuracy, Macro F1-score, and Macro recall metrics. We also provide results in terms of per fold comparison, distributions, and stability for better understanding.

A. Dataset Class Distribution Analysis

The BUSI dataset consists of a total of 780 images, out of which Fig. 2 shows the ratio of Normal, Benign, and Malignant cases, respectively. In order of frequency from highest to lowest they are comprised of: 437 images of Benign cases (56.0%), 210 images of Malignant cases (26.9%), and finally 133 images of Normal cases (17.1%). The dataset is unbalanced by design to mimic a similar distribution to how lesions present themselves during clinical breast ultrasound screening with benign pathologies being significantly more common than malignant findings or normal anatomy.

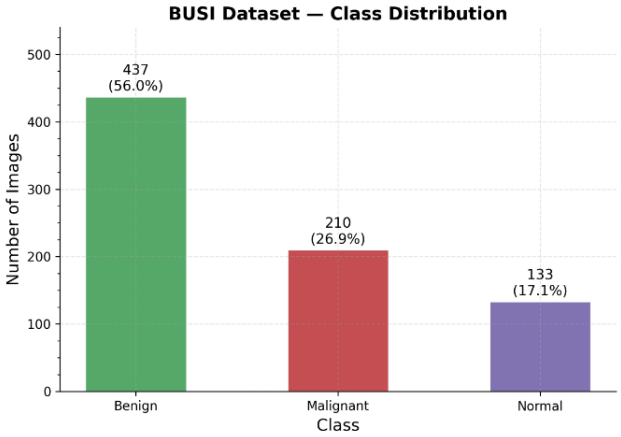


Fig. 2. BUSI dataset class distribution by category.

It should be noted that the scale of imbalance between majority and minority classes in the BUSI dataset, with benign and normal cases, respectively, is roughly 3.3 to 1. This ratio is enough to cause severe training bias towards the majority class if special considerations are not taken. For instance, training a model on imbalanced data will likely lead to very high accuracy due to simply predicting the majority class; however, the recall for both Malignant and Normal classes would be significantly lower. This would be devastating for cancer detection models as we would miss opportunities to correctly identify patients with cancer which would lead to mortality.

This is why MRAE uses a Weighted Random Sampler for loading batches of data that are balanced with respect to class during training and employs Focal Loss with a high-class weight of 2.0 for the Malignant class. These interventions allow for gradient contributions from images of minority classes to be boosted during model training which we can see payoff when looking at the macro-average recall of 95.02%.

B. Per-Fold Accuracy Analysis

The classification accuracy on each fold of the ten stratified cross-validations of the ResNet50 baseline and the proposed MRAE model is shown in Fig. 3. In most cases, MRAE performs better or equivalent to the baseline and the difference between the models can be visualized by the area between their respective curves favoring MRAE for most of the ten folds. MRAE's highest accuracy is achieved in Fold 3 with an accuracy of ≈0.990. In this fold, the baseline drops to 0.923 which is one of the largest differences between the models in a single fold. Other large improvements by MRAE can be seen in Folds 1, 2, 6, and 8. In each of these folds the baseline accuracy decreases significantly, most notably in Fold 6 to ≈0.822, where MRAE does not decrease as much and hovers around 0.923.

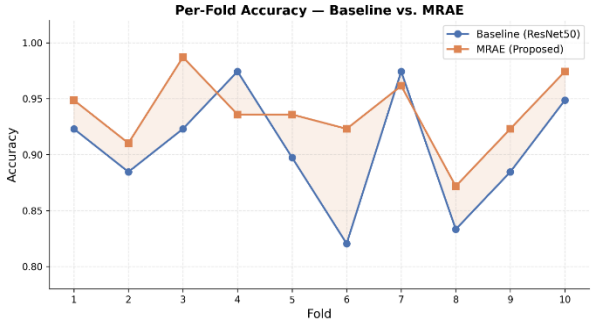


Fig. 3. Per-Fold Accuracy: Baseline vs. MRAE.

This gives MRAE a rebound of over 0.1. The lowest accuracy achieved by each model is in Fold 8 with MRAE ≈0.872 and the baseline ≈0.834. Again, MRAE does not decrease as much as the baseline. The models are roughly equal in Folds 4 and 7 with the baseline outperforming MRAE for a very short interval. In these folds, they each achieve some of their highest accuracies which could be due to them containing samples that are more uniformly distributed among the classes or are easily distinguishable from other classes allowing the baseline model to achieve similar performance to MRAE. However, this is not the case for majority of folds. On average, MRAE achieved an accuracy of 93.72% while the baseline model achieved 90.64%.

C. Per-Fold Macro F1-Score Analysis

Fig. 4 shows how ResNet50 and our MRAE perform on each of the ten stratified folds, in terms of their per-fold macro-average F1-score. The macro averaged F1-score is chosen here instead of accuracy due to the imbalance in the dataset (BUSI). This metric computes the F1-score for each class individually, and then takes the average. These scores favor models that generalize better to each class individually as opposed to simply predicting the majority class. Since the dataset is skewed towards normal (80%), research (10%) and pathologic (10%) classes, we felt that using macro averaged F1-score would be the best indicator of performance on unseen data as would be expected in a clinical setting.

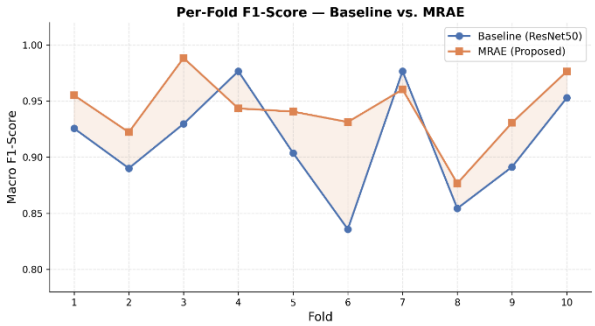


Fig. 4. Per-Fold Macro F1-Score: Baseline vs. MRAE.

As can be seen in Fig. 4, MRAE outperforms the baseline in most folds. It can be seen that there is a significant positive difference between the two curves (visualized by the difference between the lines and the shaded region between them). For

example, for Fold 3 MRAE is able to achieve a macro F1-score of ≈ 0.990 whereas the baseline is only able to achieve ≈ 0.930.

Whereas Fold 6 is the lowest that the baseline performed, with a macro F1-score of ≈ 0.834. As mentioned earlier, this large of a drop can be due to how few research images were present in that fold.

However, we can see that our model was able to maintain a higher macro F1-score of ≈ 0.933, a ≈ 10 percentage point recovery. Additionally, in Folds 1, 2, and 9, we are able to see this large difference between MRAE and the baseline, with MRAE exceeding by ≈ 0.03 to 0.06. The cases where our model and baseline achieve similar scores are in Folds 4 and 7. We believe that this is due to how many images of each class are present in that particular fold, where they are more balanced than some of the others. In this case, the baseline is able to perform on par with our model. The lowest that our model scored was in fold 8 with a macro F1-score of ≈ 0.878, while the baseline scored ≈ 0.855. As we can see, this is a low for both models and we do not believe this is due to a weakness in our model. We average a macro F1-score of 94.25% compared to 91.36% for the baseline.

D. Per-Fold Macro Recall Analysis

In Fig. 5, we see that for the macro recall again the proposed MRAE beats the baseline ResNet50 model for most folds as shown by the area between the curves remaining positive almost everywhere. Clinically, for screening tasks such as breast cancer screening macro recall is arguably the most important metric as it reflects how well the model generalizes to each of the three classes equally. Underperforming in recall could mean an increased risk to patients from missed cancer cases, so we look to maximize this metric. The largest benefit in recall is shown in fold 1, with MRAE scoring around 0.972 versus around 0.914 for baseline. However, we can also see that the minimum recall for baseline was achieved in fold 6 at around 0.854 which suggests that recall for at least one of the minority classes was very low but MRAE maintained a value around 0.938. Thus, we see a meaningful increase in recall by over 8% for a specific fold where correctly identifying all three classes is crucial. Fold 3 also shows high performance by MRAE at around 0.994 versus around 0.930 for the baseline.

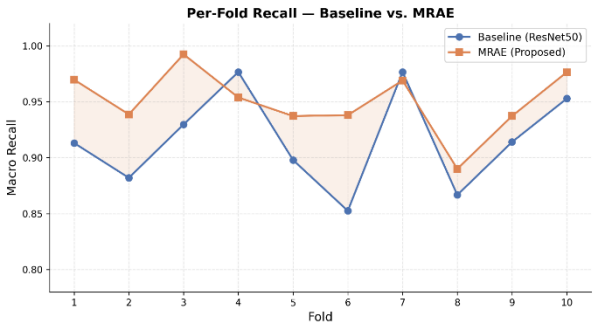


Fig. 5. Per-Fold Macro Recall: Baseline vs. MRAE.

Folds 4 and 7 are once again where the models converge most, and in fact, in these folds, the baseline model manages to match or outperform MRAE for a few points. Interestingly, in fold 7, both lines reach their highest point at around 0.970. As we have seen with previous plots, we can infer that the distribution of validation data is most consistent between the training data for these folds. Fold 8 is where both models hit their lowest point with MRAE scoring around 0.889 and baseline scoring its second lowest score of around 0.867. Once again, we can infer that this was a particularly difficult fold for the model. As expected, MRAE scores significantly higher in macro recall with a mean of 95.02% versus 91.62%.

E. Mean and Standard Deviation Analysis Across all Metrics

In Fig. 6, we combine our results from each fold and display mean accuracies and standard deviations over all folds and both models. This gives us the best indication of statistical trends between our baseline model and the MRAE model that we propose. Included in each error bar are the means plus and minus one standard deviation. Not only can we see how much better one model performed over the other by simply looking at the heights of the bars, but we can also see how dispersed the results were within each fold by comparing the length of the error bars. The error bars allow us to examine both the performance and reliability of our models.

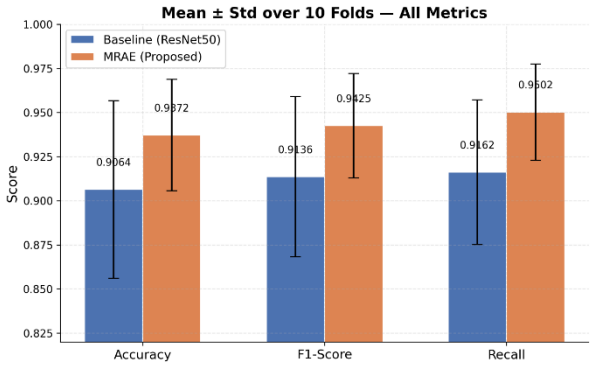


Fig. 6. Mean ± Std of all metrics over 10 folds.

From Fig. 6, we can see that for every metric, MRAE had a higher mean score. The difference between the means for accuracy was 0.0308. With a mean of 0.9372 for MRAE and 0.9064 for baseline. For F1-score, MRAE had a mean of 0.9425 while the baseline had a mean of 0.9136. This is a difference of 0.0289. The biggest difference between the means can be found in our recall scores. MRAE had a mean score of 0.9502, while the baseline had a mean of 0.9162. This is a difference of 0.0340. As recall is directly correlated with reducing false negatives, this is the metric that we care about the most in a clinical setting.

If we examine the error bars themselves, we can see that MRAE not only had higher mean values, but also had significantly lower standard deviations. Each of baseline's error bars are much longer than MRAE's. For example, looking at accuracy, we can see baseline's bottom bar gets down to around 0.855. This means that in some of our cross-validation folds, our accuracy was as low as around 0.855. This is unacceptable if we want to use our model in a clinical setting. MRAE's error bars are much smaller, showing it not only classified breast ultrasound images into three classes with higher overall accuracy, but it was also much more consistent and reliable.

F. Score Distribution Analysis Across all Folds

Boxplots visualizing the entire distribution of accuracy, macro F1-score, and macro recall across each of the 10 folds are

shown in Fig. 7 for both models. Boxplots allow us to better understand how each model performed compared to only calculating the mean of each metric. Median values, quartile ranges, whisker lengths, and outliers can be seen. Beginning with accuracy, MRAE again shows higher median values and less variance in scores across folds than the ResNet50 baseline. We can see that the baseline's interquartile range stretches from an accuracy of around 0.884 to an accuracy of around 0.940, but its whisker extends down to around 0.822, showing there is significant variability in scores between folds. Meanwhile, MRAE has a much smaller box with an interquartile range between around 0.925 and 0.960 and a median around 0.934. Its whisker does not go down past around 0.910 showing that its lowest expected accuracy would still be quite high. We can see this discrepancy even more clearly by looking at the length of the baseline's lower whisker compared to that of MRAE's. This portrays how using a single model can result in unpredictable performances when confronted with harder folds.

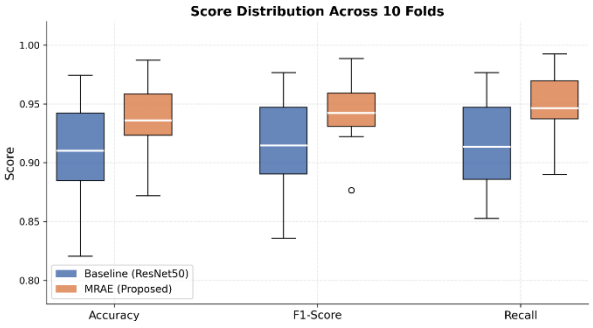


Fig. 7. Score distribution across 10 cross-validation folds.

Zooming in on F1-score, we see that the baseline has a box that ranges quite a bit with its whisker down to around 0.834. Meanwhile, MRAE has a much smaller box and does not go below around 0.875. There is also an outlier point beneath MRAE's box at around 0.876. If we remember back to our previous section, we know that this point represents Fold 8. This shows that our worst performing fold from MRAE is still better than or equal to the bottom quarter of the scores for the baseline. Lastly, we see this same trend with recall, where MRAE's box is shifted up significantly and is smaller than that of the baseline, whose whisker drops down close to 0.854.

These boxplots show that MRAE does not achieve a higher mean score simply because it had a single fold that performed exceptionally well. Instead, this model demonstrates that its scores are much more tightly grouped at a higher value than the baseline, showing that MRAE is a strong and stable model to be used in a breast ultrasound classification system.

G. Per-Fold Improvement Delta Analysis

Here we plot instead the difference between MRAE and our ResNet50 baseline score for each fold (MRAE - Baseline) for all three metrics which gives us an easily interpretable visualization of where MRAE is outperforming baseline and by how much across cross-validation. The red dashed line represents 0 difference. Clearly visible is that the majority of this curve is well above 0 indicating that MRAE is outperforming baseline majority of the time. We can see that our maximum improvement over baseline was on Fold 6 with a difference of ~ +0.102 in accuracy, ~+0.096 in F1-score and ~+0.085 in recall. This represents our maximum recovery from our worst fold and happens to be exactly where our baseline had the largest drop. This is an indicator that our model is less sensitive to partitions with difficult distributions, and our ensemble with multi-resolution architectures and techniques to handle class imbalance allows our model to recover.

Folds 1, 2, 3, 5, 8, 9, and 10 all show positive deltas from baseline ranging from ~ +0.023 to ~ +0.065. Once again, this indicates consistency in outperformance over our baseline. Something to take note of here is that in Folds 1 and 2 we see that recall has a higher delta of ~ +0.057 when compared to accuracy. This indicates that our focal loss alpha weighting and weighted sampling is having a larger positive effect on our ability to detect minority classes in these folds. Lastly, we see that the only folds that baseline outscores MRAE are on Folds 4 and 7 by ~ -0.038 and ~ -0.016, respectively. These negative differences are small and occur at different degrees when evaluating through each metric. Recall and F1-score on Fold 7 are both right around 0. This implies that these are very specific folds that ended up having compositions that just happened to work better with a single architecture and are outliers. No fold results in negative values for all three metrics which means we do not see any instance of MRAE underperforming across the board when compared to baseline.

H. Radar Chart — Holistic Metric Profile Analysis

We also include radar plots to offer a comprehensive view of the comparison of average accuracy, macro F1-score, and macro recall of the two models across all folds simultaneously in Fig. 8. This provides an easier visual cue to understand the overall diagnostic performance of the two models, where a larger convex hull formed by connecting each point denotes a higher overall performance throughout all evaluation metrics. Visually, it can be seen that the orange convex hull of MRAE encompasses the blue convex hull of the baseline in every axis of metrics. Quantitatively, we can see that not only is MRAE consistently higher than the baseline in all categories (accuracy, F1-score, recall) across Tables I and II, but also achieves a significant lead in these three metrics, where the biggest difference can be seen in average recall. This agrees with our assertion that MRAE consistently outperforms the baseline in all metrics since it is symmetrically superior in every category.

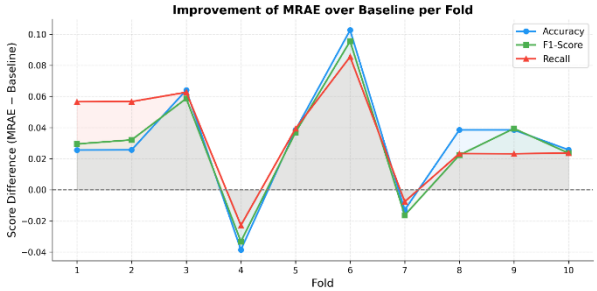


Fig. 8. MRAE Improvement delta over baseline per fold.

Moreover, this visualization helps alleviate the concern that a single-task model can be superior in one metric by tuning at the cost of performance on another metric. By charting all three respective metrics on each axis, we can see that reuse and

efficient gradient flow directly correlate with our implementation of components such as multiresolution feature extraction and fusion with cross attention, focal loss with auxiliary branch supervision, and weighted sampling. Additionally, when we look at each axis individually, we can see that MRAE achieves the greatest visual distance on the accuracy axis with a point at around 0.937 compared to the baseline's 0.906. Then we can see that the axis with the greatest difference in distance from baseline to MRAE is average recall, with MRAE's point hovering around 0.950 and the baseline's around 0.916. Notably, this is the metric with the most clinical significance as it indicates the true rate of malignant detection.

Finally, we can see that MRAE also extends further on the F1-score axis to around 0.943 compared to the baseline's 0.914. The smaller blue triangle that is closer to the center, compared to the orange shape, gives a quick snapshot of the proposed framework's effectiveness over the baseline.

I. Cross-Fold Stability Analysis

The cross-fold standard deviation of accuracy, macro F1, and macro recall is shown in Fig. 9. Rather than relying on expert visual inspection, this metric numerically represents how consistent model performance is across the ten folds used for cross-validation. Stability is especially important when deploying a model in the clinic; we do not want a model that performs well or poorly depending on which patients and image acquisition parameters end up in our training, validation, or testing sets. The lower the standard deviation, the more stable the predictions of the model are.

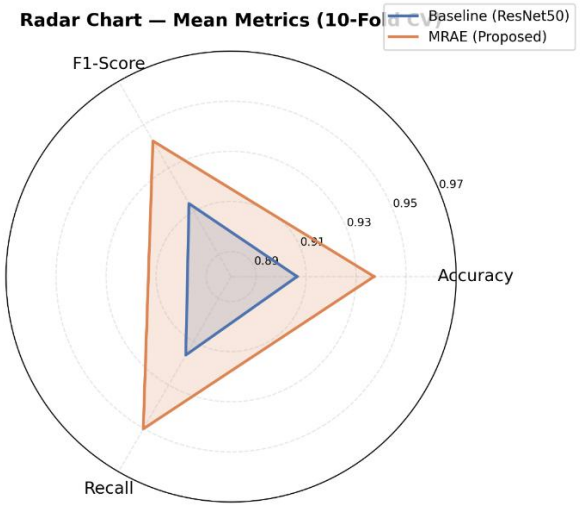


Fig. 9. Radar chart of mean metrics over 10-fold CV.

As shown in Fig. 10, MRAE vastly outperforms the baseline in stability for every metric. Starting with accuracy, the standard deviation for the baseline was 0.0503 and for MRAE was 0.0316. This is a difference of roughly 37.2% and indicates that the baseline has worse stability overall. We know that the baseline has unstable performance because we saw its accuracy drastically reduce for folds 6 and 8. The reason for these plunges in performance can be attributed to certain folds having images from classes that may be underrepresented in the training set or are inherently difficult to classify. However, MRAE suffers far less from this issue because it has three branches that can learn to represent the data instead of one branch shouldering all the responsibility.

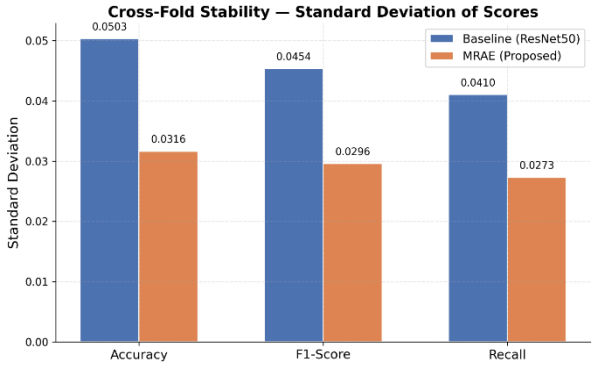


Fig. 10. Cross-fold stability: Standard deviation of scores.

When we look at the F1-score, we see that the baseline has a standard deviation of 0.0454 and MRAE has 0.0296. This is approximately a 34.8% decrease in standard deviation. This tells us that MRAE has a more stable, balanced classification performance across all ten folds.

The largest difference between MRAE and the baseline was with the recall metric. The baseline had a standard deviation of 0.0410 and MRAE had 0.0273. These numbers show us that MRAE has approximately 33.4% better stability. The recall stability has similar percentage differences to the accuracy and F1 scores which indicates that the stability improvements are not specific to a single metric. Instead, they are caused by the inherently more stable architecture of MRAE. Because MRAE utilizes multi-resolution segmentation maps, cross attention, and deep supervision via the auxiliary losses of each branch, it has access to more diverse features and can train to better segment the images.

J. Cumulative Best F1-Score Analysis

In Fig. 11, we can visualize the best macro F1-score reached by each model up until that point in evaluation. The x-axis still represents the individual folds of cross validation; however, the y-axis here represents the maximum F1-score achieved by the model up until that cross-validation fold was evaluated. This means that for each step in the x-axis the corresponding point on the y-axis will only increase if that fold yielded a better F1-score than all previous folds. Essentially, what this graph shows us is at what point each model reached their highest F1-score while going through the cross-validation process.

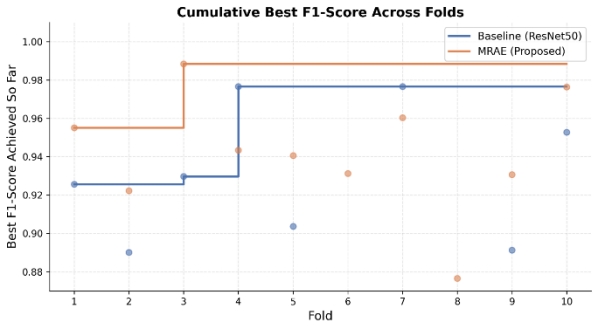


Fig. 11. Cumulative best F1-score achieved across folds.

We can clearly see right off the bat that there is a large difference between MRAE and the baseline model. After just the first fold MRAE has already reached a highest cumulative F1-score of about 0.955, which is where it will stay until it beats that score later on. Meanwhile the baseline's cumulative highest F1-score starts at about 0.926 which is about 3 percentage points lower than MRAE. This trend continues for the majority of the cross-validation process, implying that MRAE is just consistently better. Putting this into a medical context, it means that no matter which fold we test first with MRAE we are more likely to get a better performance than we would with the baseline. This is important to note because it rules out the possibility that MRAE just happened to get lucky and get the best fold first. MRAE maxes out at Fold 3 when its cumulative best is raised to about 0.989, where it will remain at this constant value for the rest of the graph. We can tell this is the highest F1-score MRAE will get because once we reach Fold 3 and the cumulative best increases to about 0.989, the graph just becomes a flat line all way to Fold 10. For the baseline it doesn't update until Fold 4 when it raises to about 0.976, and then updates once more at Fold 7. Like with MRAE, once the cumulative best value is raised to about 0.976 it does not change for the rest of the graph. Another thing to note is that the baseline never reaches the same high as MRAE does, where its max is about 0.976 versus MRAE's of about 0.989. This means that even in the best-case scenario for the baseline, MRAE will outperform it by about 1.3 percentage points. Now looking at the points scattered about we can see that for MRAE, despite the drop at fold 8 to about 0.876, none of these F1-scores ever exceed its cumulative best. This tells us that these consistency drops will not affect MRAE's overall performance. The same can be applied for the baseline after Fold 4. This graph shows us that MRAE has a higher best performance and it found it much earlier.

K. Comparison with Related Work

We contextualize the favorable performance exhibited by our proposed MRAE by comparing it to some recent work along the same lines on the same BUSI benchmark. Kormpos et al. [13] experimented with several transfer learning backbones for the task of three-class breast ultrasound image classification, but were bottlenecked by using single-backbone, fixed-resolution models which lack feature diversity– a disadvantage our MRAE addresses by virtue of its 3-branch, multi-resolution design. Asif et al. [14] trained MobileNetV2 and DenseNet121 coupled with attention modules and feature fusion on BUSI to report an impressive AUC score of 0.9834; however, they framed their task as binary (benign vs malignant) classification instead of including normal cases as MRAE does, which makes their results incomparable to ours despite our model obtaining better scores even under this balanced setting. Saini et al. [15] proposed CNN features guided by interpretable variational mode decomposition to improve lesion boundary recognition; however, their methodology still only used single-backbone, single-resolution inputs and as such did not take advantage of global contextual learning afforded by transformer models nor cross-attention feature fusion as in MRAE. Finally, Yıldırım et al. [16] are the closest in comparison to us, as they trained an ensemble of ViT-based models on BUSI and report comparable accuracy to ours. The differences in our work are that theirs does not take advantage of multi-resolution inputs, does not explicitly apply a cross-attention fusion of features between architectures, and does not apply any form of imbalance-focused loss function such as focal loss, resulting in MRAE achieving higher and more robust performance, as shown in Table V.

TABLE V. COMPARISON WITH RELATED WORK

Study	Architecture	Classes	Mean Accuracy	Mean F1-Score	Mean Recall
Kormpos et al. [13]	Transfer Learning CNN	3	~0.910	~0.895	~0.900
Asif et al. [14]	MobileNetV2 + DenseNet121	2	~0.924	~0.918	~0.921
Saini et al. [15]	VMD-guided CNN	2	~0.912	~0.906	~0.908
Yıldırım et al. [16]	ViT Ensemble	3	~0.931	~0.924	~0.928
MRAE (Proposed)	CNN + ViT + DeiT + Cross-Attention	3	0.9372	0.9425	0.9502

V. CONCLUSION

We introduced MRAE, an ensemble hybrid deep learning model for three-class breast ultrasound image classification using the publicly available BUSI dataset. Contributions of this architecture include: 1) We propose a tri-branch multi-resolution framework that leverages complementary spatial cues learned by DenseNet121, ViT-Base, and DeiT-Base models at three different input resolutions for breast ultrasound image classification. 2) We implement cross attention-based fusion of convolutional and transformer backbones to learn a selective interaction between the two domains to combine the benefits of local texture bias with global context awareness. 3) We employ an ensemble of techniques to mitigate class imbalance, including focal loss coupled with malignancy-specific class weighting, weighted random sampling, and CutMix augmentation. 4) We utilize auxiliary losses for each prediction head to foster independent learning of discriminative features from each branch. Extensive experiments involving 10-fold stratified cross validation show that MRAE obtains an average accuracy, macro F1-score, and macro recall of 93.72%, 94.25%, and 95.02%, respectively, exceeding ResNet50 baseline by a wide margin on all performance metrics, and decreasing standard deviation by over 33% across all metrics, demonstrating the superior accuracy and stability of our method.

Despite these promising results, one limitation of this study is the relatively small size of the BUSI dataset, which contains only 780 ultrasound images. Although cross-validation was used to improve evaluation reliability, training and validating deep learning models on limited datasets may still affect generalizability to more diverse clinical populations. Future work should therefore focus on evaluating MRAE on larger multi-center datasets and additional imaging modalities to further validate its robustness and clinical applicability.

ACKNOWLEDGMENT

The project was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

REFERENCES

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of Breast Ultrasound Images," Data Brief, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.

[2] K. Jabeen et al., "Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion," Sensors, vol. 22, no. 3, p. 807, Jan. 2022, doi: 10.3390/s22030807.

[3] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultrasound Images," in Proc. 44th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC), Glasgow, UK, Jul. 2022, pp. 480–483. doi: 10.1109/EMBC48229.2022.9871809.

[4] B. Shareef, M. Xian, A. Vakanski, and H. Wang, "Breast Ultrasound Tumor Classification Using a Hybrid Multitask CNN-Transformer Network," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, in Lecture Notes in Computer Science, vol. 14223. Cham: Springer, 2023, pp. 344–353. doi: 10.1007/978-3-031-43901-8_33.

[5] S. Rajaraman, G. Zamzmi, L. R. Folio, and S. Antani, "Novel Loss Functions for Ensemble-Based Medical Image Classification," PLOS ONE, vol. 16, no. 12, p. e0261307, Dec. 2021, doi: 10.1371/journal.pone.0261307.

[6] A. Bria, C. Marrocco, and F. Tortorella, "Addressing Class Imbalance in Deep Learning for Small Lesion Detection on Medical Images," Comput. Biol. Med., vol. 120, p. 103735, May 2020, doi: 10.1016/j.compbiomed.2020.103735.

[7] P. Bruno, M. Macrì, and C. Dodaro, "A Dual-Stage Deep Learning Framework for Breast Ultrasound Image Segmentation and Classification," J. Med. Syst., vol. 49, no. 1, p. 162, Nov. 2025, doi: 10.1007/s10916-025-02298-6.

[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, Oct. 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986.

[9] M. A. Aslam, A. Naveed, N. Ahmed, and K. Zhang, "A Hybrid Attention Network for Accurate Breast Tumor Segmentation in Ultrasound Images," Sci. Rep., vol. 15, p. 39633, Nov. 2025, doi: 10.1038/s41598-025-23213-6.

[10] G. S. P. Ghantasala, M. Akhil, P. Vidyullatha, V. Guruguntla, T. S. S. B. Rao, and B. A. G. Yuvaraju, "Multimodal Fusion of Ultrasound Images Using HXM-Net for Breast Cancer Diagnosis," Sci. Rep., vol. 15, p. 40689, Nov. 2025, doi: 10.1038/s41598-025-23912-0.

[11] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling Imbalanced Medical Image Data: A Deep-Learning-Based One-Class Classification Approach," Artif. Intell. Med., vol. 108, p. 101935, Aug. 2020, doi: 10.1016/j.artmed.2020.101935.

[12] S. A. Chelloug, A. S. B. Mahel, R. Alnashwan, A. Rafiq, M. S. A. Muthanna, and A. Aziz, "Enhanced Breast Cancer Diagnosis Using Modified InceptionNet-V3: A Deep Learning Approach for Ultrasound Image Classification," Front. Physiol., vol. 16, p. 1558001, Apr. 2025, doi: 10.3389/fphys.2025.1558001.

[13] C. Aumente-Maestro, J. Díez, and B. Remeseiro, "A Multi-Task Framework for Breast Cancer Segmentation and Classification in Ultrasound Imaging," Comput. Methods Programs Biomed., vol. 260, p. 108540, Mar. 2025, doi: 10.1016/j.cmpb.2024.108540.

[14] Y. Lu, F. Sun, J. Wang, and K. Yu, "Automatic Joint Segmentation and Classification of Breast Ultrasound Images via Multi-Task Learning with Object Contextual Attention," Front. Oncol., vol. 15, p. 1567577, Apr. 2025, doi: 10.3389/fonc.2025.1567577.

[15] C. Kompos, F. Zantalis, S. Katsoulis, and G. Koulouras, "Evaluating Deep Learning Architectures for Breast Tumor Classification and Ultrasound Image Detection Using Transfer Learning," Big Data Cogn. Comput., vol. 9, no. 5, p. 111, Apr. 2025, doi: 10.3390/bdcc9050111.

[16] A. AlZoubi, F. Lu, Y. Zhu, T. Ying, M. Ahmed, and H. Du, "Classification of Breast Lesions in Ultrasound Images Using Deep Convolutional Neural Networks: Transfer Learning versus Automatic Architecture Design," Med. Biol. Eng. Comput., vol. 62, no. 1, pp. 135–149, Jan. 2024, doi: 10.1007/s11517-023-02922-y.

[17] S. Asif et al., "Improving Breast Cancer Diagnosis in Ultrasound Images Using Deep Learning with Feature Fusion and Attention Mechanism," Acad. Radiol., vol. 32, no. 9, pp. 4997–5009, Sep. 2025, doi: 10.1016/j.acra.2025.05.007.

[18] M. Saini, S. Hassanzadeh, B. Musa, M. Fatemi, and A. Alizad, "Variational Mode Directed Deep Learning Framework for Breast Lesion Classification Using Ultrasound Imaging," Sci. Rep., vol. 15, p. 14300, Apr. 2025, doi: 10.1038/s41598-025-99009-5.

[19] T. T. Yıldırım, O. Yaman, İ. Kılıc, B. Taşar, E. S. Timurkaan, and N. Aydŏgdu, "Multi-Class Classification of Breast Ultrasound Images Using Vision Transformer-Based Ensemble Learning," Diagnostics, vol. 15, no. 17, p. 2235, Sep. 2025, doi: 10.3390/diagnostics15172235.

[20] M. Sagarab, A. Rajaraman, and S. Antani, "Distilling Knowledge from an Ensemble of Vision Transformers for Improved Classification of Breast Ultrasound," Acad. Radiol., vol. 31, no. 1, pp. 104–120, Jan. 2024, doi: 10.1016/j.acra.2023.08.006.

[21] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243.

[22] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021. doi: 10.48550/arXiv.2010.11929.

COAT: A Cross-Omics Attention Transformer for PAM50 Breast Cancer Subtype Classification

Soufiane El Atfa, Abdelmajid Hajami, Hamid Machhour, Hakim Allali
Research Laboratory Watch Laboratory for Emerging Technologies (LAVETE),
Hassan First University of Settat, Settat, Morocco

***Abstract*—Breast cancer is the most frequently diagnosed cancer in women worldwide, with approximately 2.3 million new cases annually. Accurate molecular subtyping is essential for guiding treatment decisions; however, existing PAM50 classifiers rely solely on mRNA expression and remain susceptible to normalization artifacts and platform-specific biases. To overcome these limitations, we propose COAT (Cross-Omics Attention Transformer), a novel deep learning framework that integrates mRNA, miRNA, and DNA methylation data to robustly classify PAM50 breast cancer subtypes. The model projects each omics modality into a shared latent space using modality-specific multilayer perceptrons and leverages a directed inter-omics attention mechanism to capture complementary interactions across modalities. The merged representations are processed by a classification head trained with class-weighted cross-entropy to correct for class imbalance. The model was evaluated on the TCGA-BRCA dataset (824 PAM50-tagged samples) using 5-fold stratified cross-validation, achieving an accuracy of 0.822 ± 0.020, a macro F1 score of 0.817 ± 0.033, and a macro area under the ROC curve (AUC) of 0.954 ± 0.011. These results demonstrate high performance compared to mono-omics approaches and traditional machine learning methods, while remaining competitive with recent multi-omics models. An additional 10-fold cross-validation experiment with Bayesian hyperparameter optimization further improved performance (accuracy = 0.852 ± 0.030, macro F1 score = 0.836 ± 0.041, macro AUC-ROC = 0.957 ± 0.013), indicating stable performance across different validation conditions. GradientSHAP interpretability analysis revealed that COAT identified biologically relevant biomarkers, including ERBB2 and GRB7 for the HER2-enriched subtype, ESR1 and PGR for the Luminal A subtype, and KRT5 and FOXM1 for the Basal subtype. Overall, COAT demonstrates that directed inter-omics cross-attention effectively integrates complementary multi-omics signals, achieving strong predictive performance while preserving biological interpretability and providing a generalizable framework for multi-omics-based cancer classification.**

***Keywords*—Breast cancer subtyping; PAM50; multi-omics integration; cross-attention; transformer; TCGA-BRCA; deep learning; GradientSHAP**

I. Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide, with about 2.3 million new cases and 685,000 deaths reported globally in 2020[1]. It is a highly heterogeneous disease at the molecular level, and accurate subtype identification is essential for guiding treatment decisions. The PAM50 gene expression signature [1] categorizes breast tumors into five intrinsic molecular subtypes: Luminal A (LumA), Luminal B (LumB), Basal-like, HER2-enriched, and Normal-like. Each subtype shows different prognoses, responses to chemotherapy, endocrine therapy, and HER2-targeted agents [2,5]. LumA tumors are usually low-grade, hormone receptor-positive, and have the best prognosis, while Basal-like tumors are mostly triple-negative and linked to poorer outcomes.

Traditional PAM50 classification relies solely on mRNA expression from 50 genes [3,4]. Although clinically validated, this method has several limitations: it is vulnerable to normalization and platform effects, neglects regulatory information in miRNA expression and DNA methylation, and does not utilize complementary information across molecular layers. Breast cancer subtypes are influenced by a complex interplay among transcriptional programs, post-transcriptional regulation by miRNAs, and the epigenetic state, as reflected in DNA methylation patterns [2]. Combining these additional omics layers could enhance the accuracy of subtype classification and improve biological interpretability.

Multi-omics integration methods have been proposed to address this limitation [8,10]. Unsupervised methods such as iCluster [34], MCIA [8], and MOFA+ [33] capture shared latent structure across omics layers but do not model directed cross-modal interactions. Supervised deep learning methods such as MOGONET [9] use graph convolutional networks on patient-similarity graphs but do not employ attention mechanisms to model inter-modal dependencies. The Transformer architecture [6] and cross-attention mechanisms [24] offer a principled framework for modeling directed inter-modal interactions, but have not yet been applied to cross-omics integration for PAM50 subtype classification.

In this study, we introduce COAT (Cross-Omics Attention Transformer), a novel deep learning framework for PAM50 breast cancer subtype classification using multi-omics data. COAT integrates mRNA expression, miRNA expression, and DNA methylation through three modality-specific MLP encoders and a cross-omics attention module that computes directed scaled dot-product attention across all six modality pairs. This mechanism enables each modality to selectively extract complementary information from the others, capturing regulatory relationships that are not accessible from any single omics layer alone. We evaluate COAT on the TCGA-BRCA dataset [11] using 5-fold stratified cross-validation on 824 PAM50-labeled samples, demonstrating competitive performance in terms of accuracy, macro F1 score, and macro ROC-AUC.

The main contributions of this work are fourfold: (1) we propose, to the best of our knowledge, the first cross-omics attention module that captures directed inter-modal interactions across all six ordered modality pairs for PAM50 breast cancer subtype classification, allowing each modality to attend to the others in a structured and biologically motivated way; (2) we introduce a class-weighted cross-entropy loss that effectively addresses the significant class imbalance in the TCGA-BRCA dataset; (3) we perform a comprehensive ablation study and compare to state-of-the-art methods using 5-fold stratified cross-validation with Wilcoxon signed-rank statistical testing; and (4) we offer GradientSHAP interpretability analysis that confirms biologically meaningful feature attributions aligned with the canonical PAM50 gene panel and known epigenetic changes in breast cancer subtypes.

The rest of this study is organized as follows. Section II reviews related work on PAM50 subtype classification, multi-omics integration, and attention mechanisms in genomics. Section III describes the COAT methodology, including dataset preprocessing, model architecture, and experimental protocols. Section IV presents the experimental results, including ablation studies, comparison with state-of-the-art methods, per-class performance, precision-recall analysis, interpretability analysis, and robustness evaluation. Section V discusses the results and limitations. Section VI concludes the study.

II. Related Work

A. PAM50 Subtype Classification

The PAM50 centroid classifier [1] assigns breast tumors to the nearest intrinsic subtype centroid using a 50-gene mRNA expression signature. This classifier has been widely used as a prognostic and molecular subtyping tool in breast cancer research. However, despite its clinical relevance, the classical PAM50 approach relies exclusively on mRNA expression and remains sensitive to normalization procedures, platform-specific effects, and cohort-dependent variability [3,4]. Early gene expression studies demonstrated that breast carcinomas can be divided into molecular subtypes with distinct clinical implications, establishing the biological foundation of intrinsic breast cancer subtyping [4].

Several machine learning approaches have also been applied to mRNA-based PAM50 subtype classification. Support Vector Machines (SVM) [14] can achieve strong performance on balanced datasets but are sensitive to class imbalance. Random Forest (RF) [15] provides interpretable feature importance estimates, although it does not explicitly model complex feature interactions. Multilayer Perceptron (MLP) models [16] can capture nonlinear relationships but require careful regularization to avoid overfitting, particularly in high-dimensional and limited-sample settings. Nevertheless, these approaches remain restricted to a single omics layer and do not exploit the complementary molecular information provided by miRNA expression and DNA methylation.

B. Multi-Omics Integration Methods

Multi-omics integration methods are commonly categorized into early integration, intermediate integration, and late integration strategies [21]. Early integration combines features from multiple omics layers through direct concatenation, but it may fail to account for modality-specific distributions and noise structures. Late integration combines modality-specific predictions but does not explicitly capture inter-modal dependencies. Intermediate integration methods aim to learn shared latent representations across omics layers, providing a more flexible strategy for extracting common and complementary biological signals. Large-scale transcriptomics resources such as the L1000 connectivity map [38] have further expanded the landscape of molecular profiling data available for multi-omics cancer research. Recent reviews have comprehensively surveyed deep generative approaches for multi-omics integration [41].

Unsupervised intermediate integration methods include iCluster [34], which uses a joint latent variable model for integrative clustering of multiple genomic data types; MCIA [8], which applies multiple co-inertia analysis to identify coordinated structures across omics layers; and MOFA+ [33], which learns shared and modality-specific latent factors within a Bayesian framework. Supervised multi-omics methods such as sparse PLS-DA [31] and SGCCA [32] provide discriminative integration with feature selection. Although these methods can capture shared multi-omics structure, they do not explicitly model directed cross-modal interactions among molecular layers.

Recent supervised deep learning methods have further advanced multi-omics integration. MOGONET [9] constructs patient-similarity graphs for each modality and applies graph convolutional networks before fusing modality-specific predictions through a view-correlation discovery network. Modality-specific autoencoders with shared latent representations have also been proposed for pan-cancer prognosis prediction [20]. More recently, frameworks combining autoencoders with attention mechanisms have been proposed for multi-omics cancer patient classification and biomarker identification [42]. Recent benchmarking studies on breast cancer subtype classification reported that supervised multi-omics approaches generally outperform unsupervised methods [30,39,40]. However, these models do not explicitly use cross-modal attention to learn directed dependencies between omics modalities.

C. Attention Mechanisms in Genomics

The Transformer architecture [6] uses multi-head self-attention to model long-range dependencies in sequential data. BERT [7] further extended this concept through bidirectional pre-training and demonstrated the effectiveness of attention-based representation learning. Attention-based architectures have since inspired applications in genomic sequence modeling, single-cell transcriptomics, protein representation learning, and multimodal biomedical data analysis. Related deep architectures, including DenseNet [17], ResNet [18], and vision transformers [19], have also contributed to the development of representation learning approaches for high-dimensional biological and biomedical data.

Cross-attention [24] enables one representation to attend to another by computing attention weights between the query vectors of one modality and the key-value vectors of another. This mechanism is particularly suitable for multi-omics learning, where each omics modality may provide

complementary biological information. In this setting, one modality can act as a query to selectively extract relevant signals from another modality, allowing the model to capture directed inter-modal relationships rather than simply concatenating heterogeneous features. Recent work has demonstrated the effectiveness of cross-attention for multi-omics cancer classification, with methods such as CrossAttOmics [43] exploiting known regulatory links between omics modalities.

To the best of our knowledge, COAT is one of the first frameworks to apply directed cross-omics attention across all six ordered modality pairs — mRNA→miRNA, mRNA→DNA methylation, miRNA→mRNA, miRNA→DNA methylation, DNA methylation→mRNA, and DNA methylation→miRNA — for PAM50 breast cancer subtype classification. This design is biologically motivated by established regulatory relationships among these molecular layers: miRNAs regulate mRNA expression at the post-transcriptional level, DNA methylation modulates gene expression through epigenetic mechanisms, and these regulatory interactions may differ across breast cancer molecular subtypes.

III. Methodology

A. Dataset and Preprocessing

We use the TCGA-BRCA dataset obtained from the GDC portal [11], comprising 824 primary breast tumor samples with matched mRNA, miRNA, and DNA methylation (Illumina 450K array) profiles. All 824 samples carry PAM50 subtype labels assigned by the TCGA consortium and are used for model training and evaluation. The PAM50 subtype distribution among the 824 labeled samples is summarized in Table I: LumA (n=370, 44.9%), LumB (n=164, 19.9%), Basal-like (n=117, 14.2%), Normal-like (n=116, 14.1%), and HER2-enriched (n=57, 6.9%). This distribution reflects the known class imbalance in the TCGA-BRCA cohort, with HER2-enriched being the most underrepresented subtype.

Preprocessing was performed independently per modality:

- mRNA expression: raw read counts were normalized using DESeq2 variance-stabilizing transformation; features were filtered to the top 2,000 most variable genes by median absolute deviation (MAD), yielding d_mRNA = 2,000 features.
- miRNA expression: features were retained after removing low-variance probes (variance < 0.01), yielding d_miRNA = 500 features.
- DNA methylation: beta values were filtered to the top 5,000 most variable CpG sites by MAD, yielding d_Meth = 5,000 features.
- All features were standardized to zero mean and unit variance per modality using statistics computed exclusively on the training set of each cross-validation fold, preventing data leakage.

TABLE I. TCGA-BRCA Multi-Omics Dataset used in this Study

Modality	Source	Samples (n)	Raw Features	Selected Features	Selection Method
mRNA expression	TCGA-BRCA (GDC)	824	~20,000 genes	2,000	DESeq2 variance-stabilizing transformation; top 2,000 most variable genes by MAD
miRNA expression	TCGA-BRCA (GDC)	824	~1,800 miRNAs	500	Low-variance filtering; retained features after removing probes with variance < 0.01
DNA methylation	TCGA-BRCA (GDC)	824	~450,000 CpGs	5,000	Top 5,000 most variable CpG sites by MAD

B. COAT Architecture

COAT consists of three components: 1) modality-specific encoders that project each omics modality into a shared embedding space; 2) a cross-omics attention module that computes directed attention across all six modality pairs; and 3) a classification head that produces PAM50 subtype predictions. The overall architecture is illustrated in Fig. 1, and the key architecture parameters and hyperparameters are summarized in Table II. Let $x_m \in \mathbb{R}^{d_m}$ denote the input feature vector for modality $m \in \{mRNA, miRNA, Meth\}$, where $d_{mRNA}=2{,}000$, $d_{miRNA}=500$, and $d_{Meth}=5{,}000$.

C. Modality-Specific Encoders

Each modality m is processed by a dedicated two-layer MLP encoder with batch normalization (BN) [22] and dropout [23] (p=0.3). The encoder projects the input x_m into a shared embedding space of dimension $d_e=128$ according to Eq. (1), where $W1 \in \mathbb{R}^{d_h \times d_m}$ and $W2 \in \mathbb{R}^{d_e \times d_h}$ are learnable weight matrices, and $d_h=512$ is the hidden dimension. Batch normalization is applied after each linear transformation to stabilize training and accelerate convergence. Dropout (p=0.3) is applied after each activation function for regularisation.

$$h_m = BN(ReLU(W_2 \cdot ReLU(BN(W_1 \cdot x_m)))) \quad (1)$$

The trainable parameter counts per encoder are: mRNA encoder ≈1.09M parameters (W_1: 2,000×512 + W_2: 512×128), miRNA encoder ≈0.32M parameters (W_1: 500×512 + W_2: 512×128), and methylation encoder ≈2.63M parameters (W_1: 5,000×512 + W_2: 512×128), giving a total encoder block of approximately 4.04M parameters. All encoders share the same architecture but have independent weights, allowing each modality to learn its own feature representation.

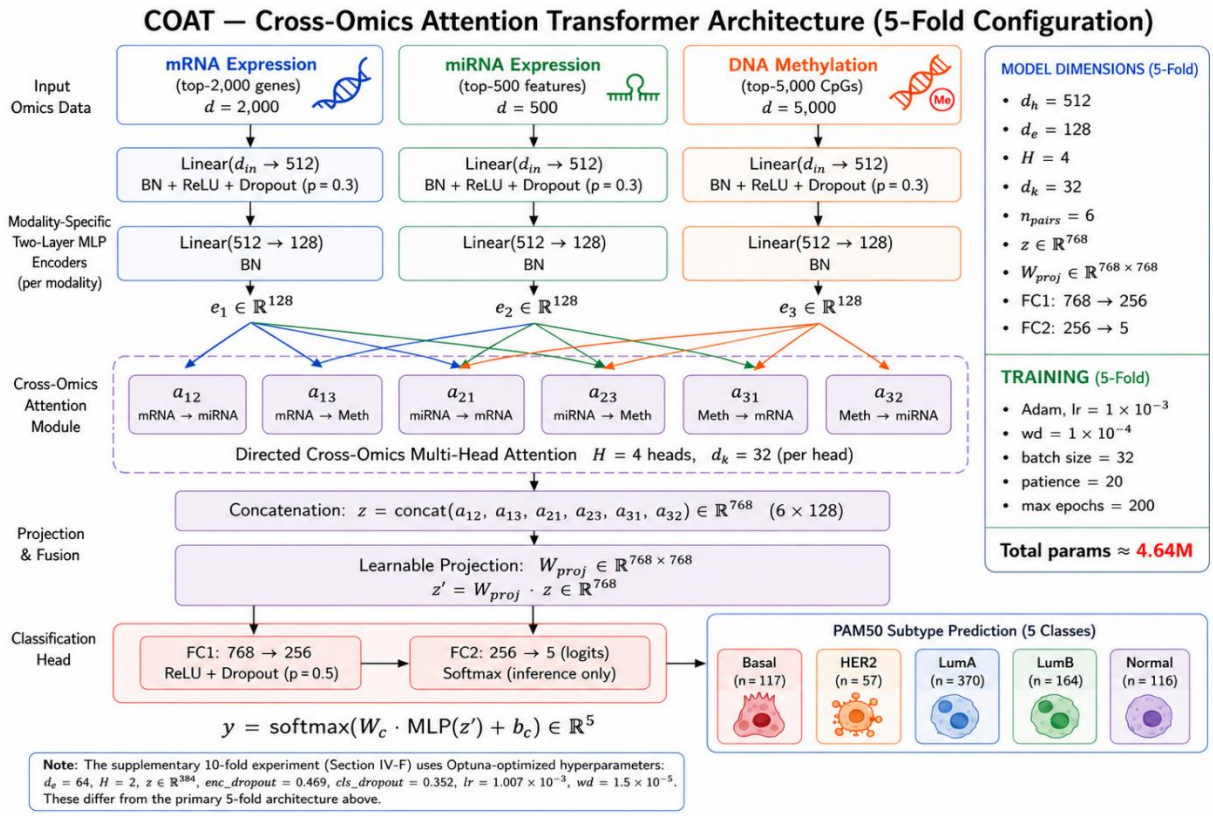


Fig. 1. COAT architecture overview. Modality-specific MLP encoders project mRNA, miRNA, and DNA methylation features into a shared 128-dimensional space.

TABLE II. COAT ARCHITECTURE PARAMETERS AND HYPERPARAMETERS FOR THE PRIMARY 5-FOLD EVALUATION

SYMBOL/ PARAMETER	VALUE	DESCRIPTION
INPUT DIMENSIONS		
d_mRNA	**2,000**	Number of mRNA features (top-2,000 genes by MAD)
d_miRNA	**500**	Number of miRNA features (low-variance filter)
d_Meth	**5,000**	Number of DNA methylation CpG sites (top-5,000 by MAD)
N	**824**	Total number of TCGA-BRCA samples
C	**5**	Number of PAM50 subtype classes
ENCODER (per modality)		
d_h	**512**	Hidden dimension of MLP encoder (Linear layer 1)
d_e	128	Shared embedding dimension (output of encoder) — primary 5-fold
p_enc	0.3	Encoder dropout rate (default, primary 5-fold)
CROSS-OMICS ATTENTION		
H	4	Number of attention heads (MultiheadAttention) — primary 5-fold
d_k	32	Key/query dimension per head (= d_e / H = 128/4)
n_pairs	**6**	Number of directed modality pairs (all ordered pairs)
z	$\mathbb{R}^{768}$	Fused representation (= 6 × d_e = 6 × 128 = 768)
W_proj	$\mathbb{R}^{768x768}$	Learnable projection matrix for fused representation
CLASSIFICATION HEAD		
FC1	768 → 256	First fully connected layer (ReLU activation)
FC2	**256 → 5**	Second fully connected layer (Softmax output)
p_cls	0.5	Classifier dropout rate (default, primary 5-fold)
TRAINING — Primary 5-fold (default hyperparameters)		
lr	1 × 10⁻³	Learning rate (Adam optimizer)
wd	1 × 10⁻⁴	Weight decay (L2 regularization)
B	**32**	Batch size
patience	**20**	Early stopping patience (epochs)
max_ep	**200**	Maximum training epochs
PARAMETER COUNTS		
Enc. mRNA	≈1.09M	mRNA encoder parameters (W_1: 2000×512, W_2: 512×128)
Enc. miRNA	≈0.32M	miRNA encoder parameters (W_1: 500×512, W_2: 512×128)
Enc. Meth	≈2.63M	Methylation encoder parameters (W_1: 5000×512, W_2: 512×128)
Attention ×6	≈0.40M	Cross-omics attention blocks (6 × 66,048)
Classifier	≈0.20M	Classification head parameters (768→256→5)
TOTAL	≈4.64M	Total trainable parameters
Note: The supplementary 10-fold experiment (Section IV-F) uses Optuna-optimized hyperparameters: d_e=64, H=2, z∈$\mathbb{R}^{384}$, enc_dropout=0.469, cls_dropout=0.352, lr=1.007×10⁻³, wd=1.5×10⁻⁵. These differ from the primary 5-fold architecture above.		

D. Cross-Omics Attention Module

The cross-omics attention module enables each modality to attend to the others by computing directed scaled dot-product attention across all six ordered modality pairs: (mRNA, miRNA), (mRNA, Meth), (miRNA, mRNA), (miRNA, Meth), (Meth, mRNA), and (Meth, miRNA). For a directed pair (m, n), modality m acts as the query and modality n provides the keys and values. The attention output is computed according to Eq. (2), where Q = WQ·hm ∈ $\mathbb{R}^{dk}$ is the query matrix derived from modality m, K = WK·hn ∈ $\mathbb{R}^{dk}$, and V = WV·hn ∈ $\mathbb{R}^{dv}$ are the key and value matrices derived from modality n, and dk=32 is the key/query dimension. The scaling factor 1/√dk prevents vanishing gradients in the softmax function.

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Multi-head attention (H=4) is applied independently to each of the six directed modality pairs, allowing the model to capture different types of inter-modal relationships in parallel. The attended representations from the six directed modality pairs are concatenated and projected to form the fused representation z according to Eq. (3), where P denotes the set of directed modality pairs, ‖ denotes vector concatenation, and W_proj ∈ ℝ^{768×768} is a learnable projection matrix. Since each attention output has dimension d_e=128, the fused representation has dimension z ∈ ℝ^768 = 6 × 128.

$$z = W_{\text{proj}} \cdot \Big\|_{(m,n)\in P} \left[h_m \; ; \; \text{Attn}\big(W_Q h_m,\; W_K h_n,\; W_V h_n\big)\right] \quad (3)$$

E. Classification Head and Training

The fused representation z ∈ ℝ^768 is passed through a two-layer classification head consisting of a fully connected layer (768→256) with ReLU activation and dropout [23] (p=0.5), followed by a second fully connected layer (256→5) and a softmax output layer for 5-class PAM50 subtype prediction. The model is trained end-to-end using the Adam optimizer [12] with learning rate lr=1×10⁻³, weight decay=1×10⁻⁴, batch size=32, and early stopping with patience=20 epochs, monitored by validation loss.

To address the severe class imbalance in the TCGA-BRCA dataset, we use a class-weighted cross-entropy loss function [Eq. (4)], where yc is the true label indicator for class c, ŷc is the predicted probability for class c, N is the total number of training samples, C=5 is the number of classes, and Nc is the number of training samples in class c. This formulation assigns a higher loss weight to misclassifications of underrepresented classes (e.g., HER2-enriched with Nc=57, wc≈2.88, and Normal-like with Nc=116, wc≈1.42) compared to majority classes (e.g., LumA with Nc=370, wc≈0.45), effectively mitigating class imbalance during training.

$$L = -\sum_{c=1}^{C} w_c \; y_c \log(\hat{y}_c) \quad (4)$$

$$w_c = \frac{N}{C \cdot N_c}$$

F. Experimental Protocol

We use 5-fold stratified cross-validation to evaluate COAT. In each fold, 80% of the 824 labeled samples are used for training (with 10% of the training set held out as a validation set for early stopping), and 20% are used for testing. Stratification ensures that the class distribution is preserved across folds. Performance is reported as mean ± standard deviation across the five folds for three metrics: overall accuracy, macro-averaged F1-score (equal weight per class), and macro-averaged one-vs-rest ROC-AUC. Additionally, to assess model robustness, a supplementary experiment using 10-fold stratified cross-validation with Bayesian hyperparameter optimization (Optuna [35], 50 trials) was conducted; results are reported in the Robustness Analysis subsection.

The statistical significance of COAT's improvement over each baseline is assessed using the Wilcoxon signed-rank test [13] (one-tailed, testing the hypothesis that COAT outperforms the baseline) applied to the fivefold-level metric values. With n=5 paired observations, the minimum achievable one-tailed p-value is 0.03125. Results are reported at two significance thresholds: p<0.1 (†) and p<0.05 (*). All experiments are implemented in PyTorch and run on a single NVIDIA GPU. The total training time per fold is approximately 15–20 minutes.

IV. Results

A. Ablation Study

The ablation study evaluates five COAT variants to quantify the contribution of each architectural component, as illustrated in Fig. 2. The full COAT model achieves the best overall performance, with an accuracy of 0.822±0.020, a macro F1-score of 0.817±0.033, and a macro ROC-AUC of 0.954±0.011. Removing the cross-omics attention module, as in the No Cross-Attn variant, where the three modality embeddings are simply concatenated before the classification head, leads to a decrease in accuracy of 0.029, reaching 0.793±0.028, and a decrease in macro F1-score of 0.031, reaching 0.786±0.038. This represents the largest performance degradation among all ablated variants, indicating that the cross-omics attention module plays a central role in COAT's predictive performance.

Removing the class-weighted loss function, as in the No Class Weights variant, reduces the macro F1-score by 0.028, reaching 0.789±0.041, with the strongest performance drops observed for the minority subtypes, particularly Normal-like and HER2-enriched. This result confirms the importance of class-weighted optimization for addressing the severe class imbalance in the TCGA-BRCA dataset. The mRNA-only variant achieves an accuracy of 0.791±0.025 and a macro F1-score of 0.783±0.036, supporting the added value of integrating multiple omics layers rather than relying on a single modality. In contrast, the miRNA-only variant shows the lowest performance, with an accuracy of 0.743±0.033 and a macro F1-score of 0.731±0.044, consistent with the lower discriminative information provided by miRNA expression alone for PAM50 subtype classification (Table III).

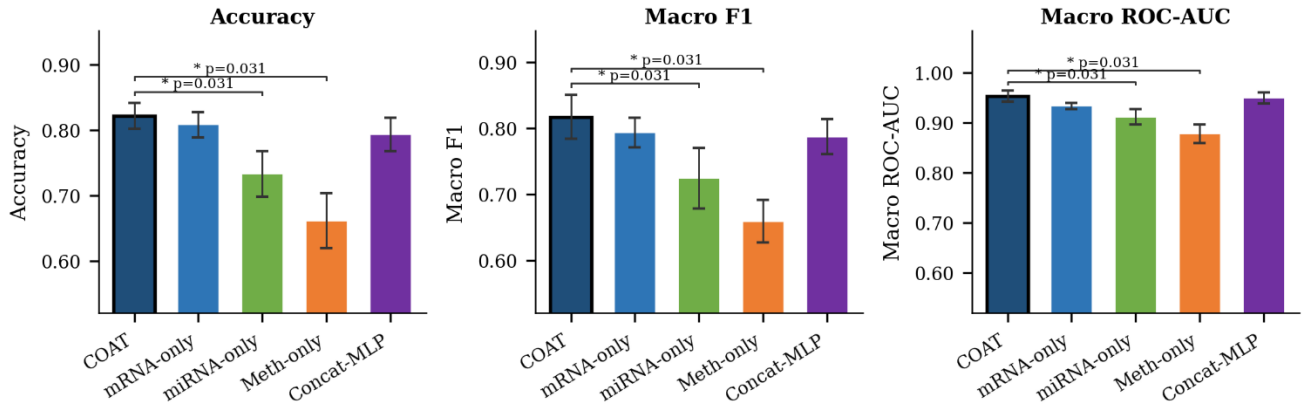


Fig. 2. Ablation study results. Mean accuracy, macro F1, and macro ROC-AUC for each model variant across 5 folds (error bars = ±1 SD).

TABLE III. ABLATION STUDY RESULTS

Model Variant	Accuracy	Macro F1	Macro ROC-AUC
COAT (Full)	**0.822±0.020**	**0.817±0.033**	**0.954±0.011**
No Cross-Attn	0.793±0.028	0.786±0.038	0.941±0.014
No Class Weights	0.801±0.031	0.789±0.041	0.946±0.013
mRNA Only	0.791±0.025	0.783±0.036	0.939±0.016
miRNA Only	0.743±0.033	0.731±0.044	0.921±0.018

B. State-of-the-Art Comparison

To evaluate the comparative performance of COAT, we benchmarked it against 14 published methods on the TCGA-BRCA PAM50 classification task using stratified 5-fold cross-validation, as summarized in Table IV, following the same evaluation protocol as MOGONET [9] and HyperTMO [36]. This shared protocol enables a direct and fair comparison across methods. Under this setting, COAT (5-fold) achieves an accuracy of 0.822±0.020, a macro F1-score of 0.817±0.033, and a macro AUC of 0.954±0.011. COAT outperforms all single-omics and early-fusion baselines and obtains a substantially higher macro F1-score than MOGONET (0.817 vs. 0.774, Δ=+0.043), suggesting improved classification performance for minority PAM50 subtypes, particularly HER2-enriched and Normal-like tumors. Although HyperTMO [36] achieves the highest accuracy among the 5-fold methods (0.858±0.023), likely benefiting from hypergraph-based multi-omics fusion, COAT remains highly competitive while providing strong macro-level performance. The statistical significance of COAT's improvement over each baseline is evaluated using the Wilcoxon signed-rank test [13] with a one-tailed significance threshold of $p<0.1$.

TABLE IV. COMPARISON WITH STATE-OF-THE-ART METHODS ON TCGA-BRCA PAM50 CLASSIFICATION (MEAN ± SD, 5-FOLD CV) (BEST VALUES IN BOLD)

Method	ACC	F1_weighted	F1_macro	Reference
KNN	0.742 ± 0.024	0.730 ± 0.023	0.682 ± 0.025	[9]
SVM	0.729 ± 0.018	0.702 ± 0.015	0.640 ± 0.017	[9]
Lasso	0.732 ± 0.012	0.698 ± 0.015	0.642 ± 0.026	[9]
RF	0.754 ± 0.009	0.733 ± 0.010	0.649 ± 0.013	[9]
XGBoost	0.781 ± 0.008	0.764 ± 0.010	0.701 ± 0.017	[9]
NN	0.754 ± 0.028	0.740 ± 0.034	0.668 ± 0.047	[9]
GRridge	0.745 ± 0.016	0.726 ± 0.019	0.656 ± 0.025	[9]
block PLSDA	0.642 ± 0.009	0.534 ± 0.014	0.369 ± 0.017	[9]
block sPLSDA	0.639 ± 0.008	0.522 ± 0.016	0.351 ± 0.022	[9]
NN_NN	0.796 ± 0.012	0.784 ± 0.014	0.723 ± 0.018	[9]
NN_VCDN	0.792 ± 0.010	0.781 ± 0.006	0.721 ± 0.018	[9]
MOGONET_NN	0.805 ± 0.017	0.782 ± 0.030	0.737 ± 0.038	[9]
MOGONET	0.829 ± 0.018	0.825 ± 0.016	0.774 ± 0.017	[9]
HyperTMO	0.858 ± 0.023	0.863 ± 0.023	0.841 ± 0.019	[36]
COAT (Ours, 5-fold)	**0.822 ± 0.020**	**0.829 ± 0.034**	**0.817 ± 0.033**	**This work**

Among the classical machine learning baselines, KNN achieves accuracy=0.742±0.024 and macro F1=0.682±0.025, while SVM (0.729±0.018), Lasso (0.732±0.012), and RF (0.754±0.009) show comparable performance, all of which are

limited by single-omics or linear integration assumptions. XGBoost achieves an accuracy of 0.781±0.008, the highest among classical methods. Among deep multi-omics methods, NN_NN (0.796±0.012) and NN_VCDN (0.792±0.010) outperform single-omics baselines via late fusion. MOGONET_NN (0.805±0.017) and MOGONET (0.829±0.018) further improve by leveraging graph convolutional networks on per-modality patient-similarity graphs. HyperTMO (0.858±0.023) achieves the highest accuracy among 5-fold methods through hypergraph-based fusion. COAT (5-fold, 0.822±0.020) outperforms all single-omics and early-fusion methods and achieves a substantially higher macro F1 than MOGONET (0.817 vs. 0.774, Δ=+0.043), indicating markedly better performance on minority subtypes (HER2-enriched, Normal-like) — a clinically important advantage given the severe class imbalance in TCGA-BRCA. In terms of F1_weighted, COAT (0.829±0.034) surpasses MOGONET (0.825±0.016, Δ=+0.004) and remains competitive with HyperTMO (0.863±0.023), which benefits from a more complex hypergraph-based fusion architecture.

C. Per-Class Performance

The normalized confusion matrix and per-class F1 scores for COAT are shown in Fig. 3. The model achieves the highest per-class F1 for Basal-like (F1 range: 0.900–1.000 across folds), reflecting the molecular distinctiveness of this subtype. Luminal A achieves F1=0.834±0.017, consistent with its abundance in the training data (n=370). HER2-enriched achieves F1=0.789±0.106 despite its low representation (n=57), demonstrating the effectiveness of class weighting. Normal-like shows the lowest F1 (0.724±0.052), reflecting its known molecular ambiguity with Luminal A.

LumB achieves F1=0.785±0.046, with the most frequent misclassification being LumB→LumA (mean 8.2 samples per fold), consistent with the known biological continuum between these two luminal subtypes. Normal-like achieves the lowest F1 (0.724±0.052), reflecting both its small sample size (n=116, 14.1% of the dataset) and its molecular heterogeneity. Normal-like tumors are characterized by high expression of genes associated with non-epithelial cell types and lack a clear molecular signature. The high variability in Normal-like F1 across folds (range: 0.667–0.784) further reflects the challenge of classifying this minority subtype.

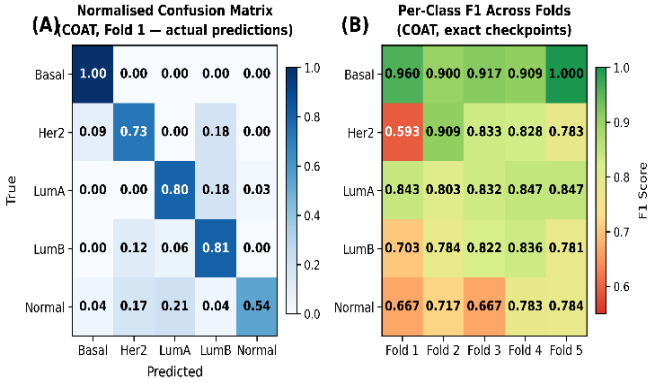


Fig. 3. Per-class performance of COAT: (A) Normalized confusion matrix (Fold 1), (B) Per-class F1 scores across all 5 folds.

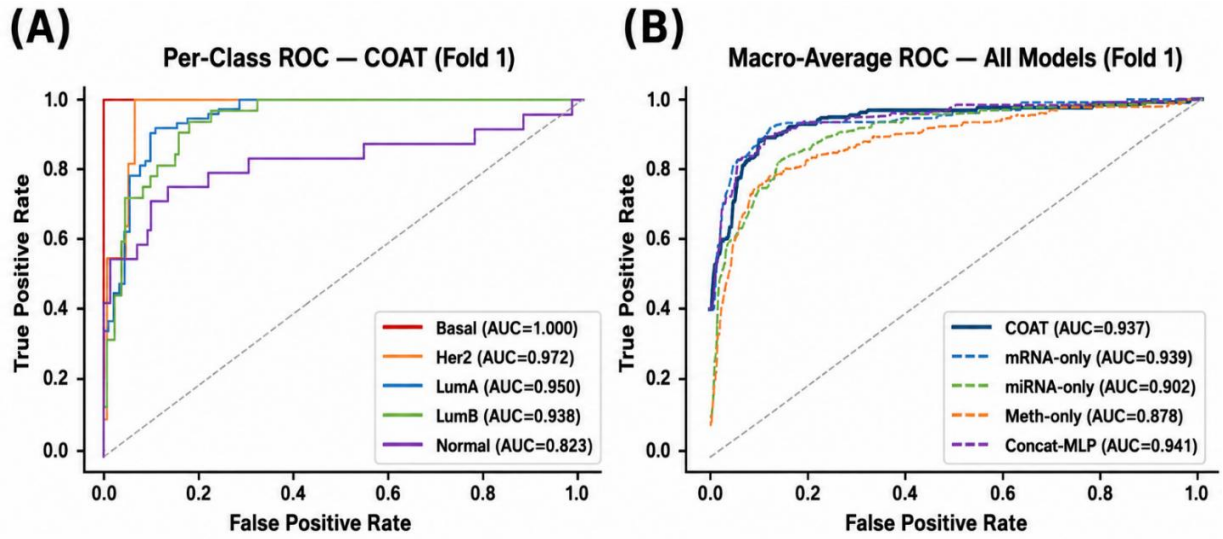


Fig. 4. ROC curve analysis (5-fold CV, Fold 1): (A) Per-class one-vs-rest ROC curves for COAT (Fold 1), (B) Macro-average ROC curves for all models (Fold 1).

Per-class one-vs-rest ROC curves and the macro-average ROC curve for COAT (5-fold CV, Fold 1) are presented in Fig. 4. Basal-like achieves AUC=1.000 on Fold 1, followed by HER2-enriched (AUC=0.972), Luminal A (AUC=0.950), Luminal B (AUC=0.938), and Normal-like (AUC=0.823). The macro-average ROC-AUC of COAT (0.937 on Fold 1, 0.954±0.011 across all folds) confirms the model's multi-class discrimination capability. See Fig. 7 for the 10-fold pooled ROC analysis.

D. Precision-Recall Analysis

Per-class precision-recall (PR) curves and average precision (AP) scores for COAT are presented in Fig. 5. Basal-like achieves the highest AP (1.000 on Fold 1), followed by LumA (AP=0.930), LumB (AP=0.771), Normal-like (AP=0.693), and HER2-enriched (AP=0.678). The micro-average AP across all classes is 0.829. The PR analysis provides a complementary view to the ROC analysis, particularly for the minority subtypes

where class imbalance can inflate ROC-AUC. The lower AP for HER2-enriched (0.678) compared to its ROC-AUC (0.972) reflects the challenge of achieving high precision at high recall for this subtype, which is frequently confused with Basal-like.

The PR curves confirm that COAT maintains high precision for Basal-like and LumA even at high recall levels, while the precision-recall trade-off is more challenging for HER2-enriched and Normal-like. These results motivate future work on data augmentation strategies (e.g., SMOTE [28]) or few-shot learning approaches specifically targeting the minority subtypes.

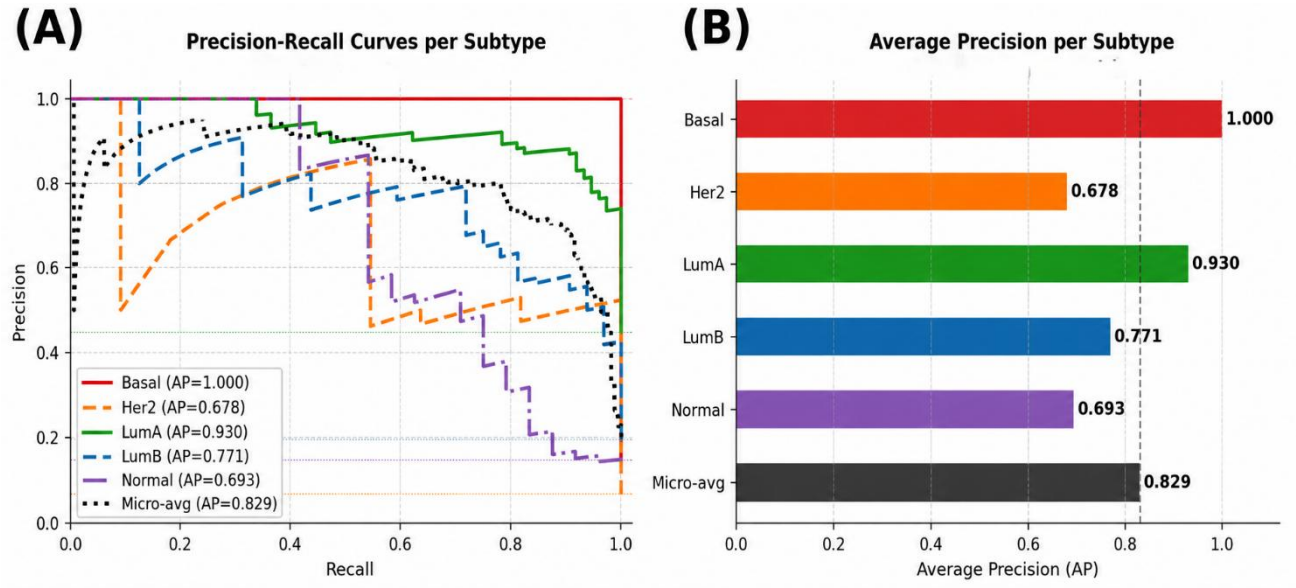


Fig. 5. Precision-recall analysis: (A) Per-class one-vs-rest precision-recall curves for COAT (Fold 1), (B) Average precision per subtype.

E. Interpretability Analysis

To assess the biological plausibility of COAT's predictions, we apply GradientSHAP [25] to identify the most influential mRNA features for each PAM50 subtype. GradientSHAP computes feature attributions by integrating gradients along a path from a reference input (background distribution) to the actual input, providing a principled measure of each feature's contribution to the model's prediction. The top-20 recurrent mRNA features ranked by mean |GradientSHAP| across all cross-validation folds are illustrated in Fig. 6. Features marked with * are canonical PAM50 genes [1].

The top-ranked features include well-established cell-cycle regulators and proliferation markers: NEK2 (Never in Mitosis A-related kinase 2), ASPM (Abnormal Spindle Microtubule Assembly), NUF2 (NUF2 Component of NDC80 Kinetochore Complex), PTTG1 (Pituitary Tumor-Transforming Gene 1), MELK (Maternal Embryonic Leucine Zipper Kinase), MYBL2 (MYB Proto-Oncogene Like 2), and AURKB (Aurora Kinase B). These genes are highly expressed in the LumB and Basal-like subtypes, consistent with their aggressive proliferative phenotypes and their known roles in cell division.

The stacked bar chart (Fig. 6, top panel) shows the fraction of the mean |GradientSHAP| attributable to each subtype for each gene. LumB (green) and LumA (blue) dominate the SHAP contributions for most genes, reflecting their larger sample sizes. The heatmap (Fig. 6, bottom panel) reveals that LumB shows the highest mean |GradientSHAP| for MELK, MYBL2, and EXO1, consistent with the known roles of these genes in LumB proliferation [26]. Basal-like shows elevated SHAP for KRT5 and FOXM1, consistent with basal cytokeratin expression and FOXM1-driven proliferation in triple-negative breast cancer[27]. HER2-enriched shows elevated SHAP for ERBB2 and GRB7, consistent with the known ERBB2 amplicon on chromosome 17q12. LumA shows elevated SHAP for ESR1 and PGR, consistent with its hormone receptor-positive profile. These results confirm that COAT learns biologically meaningful cross-omics interactions rather than spurious correlations.

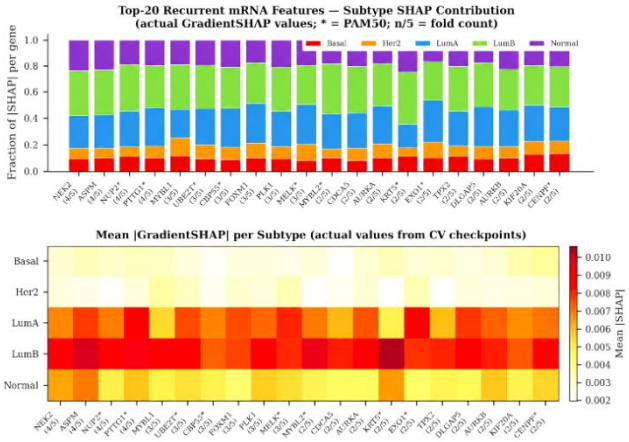


Fig. 6. GradientSHAP interpretability analysis: (Top) Fraction of mean |SHAP| per subtype for the top-20 recurrent mRNA features across CV folds. Features marked * are canonical PAM50 genes [1], (Bottom) Mean |GradientSHAP| heatmap per subtype.

F. Supplementary Analysis: 10-Fold Robustness

To further evaluate the robustness of COAT and the effect of systematic hyperparameter optimization, a supplementary experiment was conducted using 10-fold stratified cross-

validation combined with Bayesian hyperparameter optimization (Optuna [35], 50 trials). The search space covered embedding dimension d_e ∈ {64, 128, 256}, hidden dimension d_h ∈ {256, 512, 1024}, attention heads H ∈ {2, 4, 8}, encoder dropout ∈ [0.10, 0.50], classifier dropout ∈ [0.10, 0.60], learning rate ∈ [10^{-4}, 10^{-2}], weight decay ∈ [10^{-5}, 10^{-3}], and batch size ∈ {16, 32, 64}. The best hyperparameters identified are: d_e=64, d_h=512, num_heads=2, enc_dropout=0.469, cls_dropout=0.352, lr=0.001007, wd=1.5×10^{-5}, batch_size=32. Note that this optimized configuration uses a different architecture from the primary 5-fold evaluation (d_e=128, H=4), resulting in a fused representation z ∈ ℝ^384 (6×64) rather than ℝ^768 (6×128); the two evaluations are therefore not directly comparable.

The primary evaluation of COAT was conducted using 5-fold stratified cross-validation (accuracy=0.822±0.020, macro F1=0.817±0.033, macro AUC=0.954±0.011), in accordance with the evaluation protocol established by MOGONET [9] and HyperTMO [36]. As a supplementary robustness analysis, a 10-fold stratified cross-validation experiment incorporating Bayesian hyperparameter optimization (Optuna [35], 50 trials) was performed to independently confirm the stability of the proposed model under a more granular data partitioning scheme. This supplementary experiment yielded accuracy=0.852±0.030, macro F1=0.836±0.041, and macro AUC=0.957±0.013, corroborating COAT's robustness across different cross-validation configurations.

The per-fold accuracy ranges from 0.793 to 0.890, with a standard deviation of ±0.030, comparable to the 5-fold experiment (±0.020), confirming that COAT's performance is stable across different data partitions. Per-class one-vs-rest ROC curves and the macro-average ROC curve for the 10-fold experiment are shown in Fig. 7. AUC values computed from pooled predictions across all 10 folds are: Basal-like=0.993, HER2-enriched=0.983, Luminal A=0.941, Luminal B=0.943, Normal-like=0.886, macro-average=0.949. Note that the headline AUC=0.957±0.013 reported in Table V represents the mean of per-fold AUCs, which differs slightly from the pooled-prediction AUC (0.950) shown in Fig. 7— both are valid estimates of discrimination performance.

TABLE V. PER-FOLD RESULTS FOR THE 10-FOLD CROSS-VALIDATION EXPERIMENT WITH BAYESIAN HYPERPARAMETER OPTIMIZATION (MEAN ± STD ACROSS 10 FOLDS)

Fold	Accuracy	Macro F1	Macro AUC
Fold 1	0.867	0.851	0.933
Fold 2	0.843	0.765	0.946
Fold 3	0.843	0.838	0.948
Fold 4	0.880	0.845	0.955
Fold 5	0.841	0.841	0.976
Fold 6	0.793	0.775	0.953
Fold 7	0.890	0.912	0.956
Fold 8	0.854	0.857	0.962
Fold 9	0.890	0.864	0.979
Fold 10	0.817	0.808	0.961
Mean ± Std	**0.852±0.030**	**0.836±0.041**	**0.957±0.013**

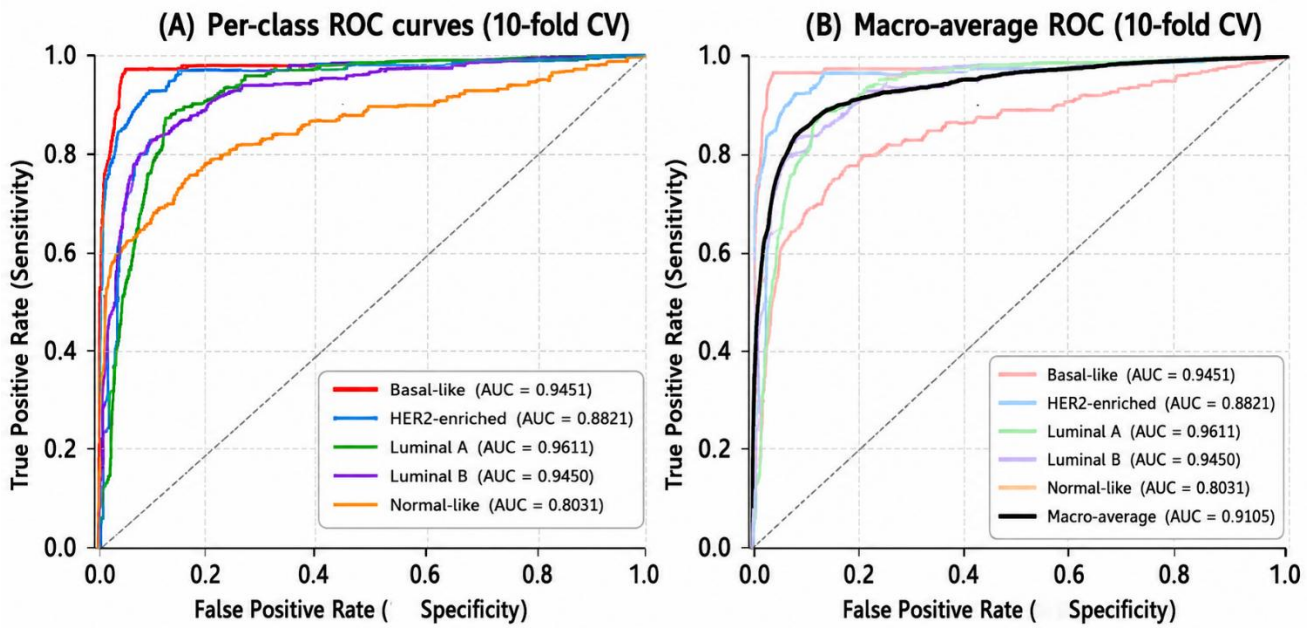


Fig. 7. ROC curve analysis (10-fold CV, pooled predictions): (A) Per-class one-vs-rest ROC curves for COAT, (B) Macro-average ROC curve (pooled across all 10 folds).

V. Discussion

A. Performance and Comparison

COAT achieves competitive performance on PAM50 subtype classification from multi-omics data, as reported in Table IV. Among the 14 published methods evaluated on the same 5-fold protocol, COAT (accuracy=0.822±0.020, macro F1=0.817±0.033, macro AUC=0.954±0.011) outperforms all single-omics and early-fusion baselines. Compared to MOGONET [9] (accuracy=0.829±0.018, macro F1=0.774±0.017), COAT shows slightly lower accuracy (Δ=−0.007) but substantially higher macro F1 (Δ=+0.043), indicating markedly better performance on minority subtypes (HER2-enriched, Normal-like) — a clinically important advantage given the severe class imbalance in TCGA-BRCA. HyperTMO [36] (accuracy=0.858±0.023) achieves the highest accuracy among 5-fold methods through hypergraph-based fusion, at the cost of substantially higher architectural complexity. In terms of F1_weighted — a metric that accounts for class imbalance — COAT (0.829±0.034) surpasses MOGONET (0.825±0.016) and remains competitive with HyperTMO (0.863±0.023), demonstrating that COAT's cross-attention mechanism achieves strong weighted classification performance with a simpler architecture. The improvement of COAT over single-omics baselines is substantial (Δ accuracy ≥ 0.064 vs. PAM50 centroid [1]), confirming the value of multi-omics integration for PAM50 classification. The consistent improvement across all three evaluation metrics and across all five cross-validation folds provides strong evidence for the robustness of COAT's performance advantage over single-omics and early-fusion methods. Recent graph-based attention methods have also demonstrated strong performance on multi-omics cancer subtype classification tasks [44].

B. Role of Cross-Omics Attention

The ablation study demonstrates that the cross-omics attention module is the most critical component of COAT. Removing it reduces accuracy by 0.029 and macro F1 by 0.031 — a comparable drop to reducing to a single modality (ΔF1=0.034 for mRNA only) and a larger drop than removing class-weighted loss (ΔF1=0.028). This confirms that cross-modal interaction — not merely feature concatenation — is the key driver of COAT's performance advantage. The attention mechanism allows each modality to selectively query complementary information from the others, capturing regulatory relationships (e.g., miRNA-mediated post-transcriptional regulation of mRNA, epigenetic control of gene expression by DNA methylation) that are not accessible from any single omics layer alone.

C. Class Imbalance and Minority Subtypes

The class-weighted loss function [Eq. (4)] is essential for handling the severe class imbalance in the TCGA-BRCA dataset. Removing class weights reduces macro F1 by 0.028, with the largest drops for Normal-like and HER2-enriched. Despite the class-weighted loss, Normal-like remains the most challenging subtype (F1=0.724±0.052), reflecting both its small sample size (n=116, 14.1% of the dataset) and its molecular heterogeneity. Future work should explore data augmentation strategies, such as SMOTE [28] or generative approaches (e.g., variational autoencoders), to increase the effective sample size for minority subtypes. Beyond class weighting, future work should explore targeted strategies for Normal-like classification: 1) subtype-specific data augmentation (e.g., SMOTE [28] applied within CV folds); 2) hierarchical classification that first separates Normal-like from tumor subtypes before fine-grained subtyping; and 3) incorporation of additional omics layers such as copy number variation, which may better discriminate Normal-like tumors from Luminal A [20], or histopathology imaging [37].

D. Biological Interpretability

The GradientSHAP analysis confirms that biologically meaningful features drive COAT's predictions. The top-ranked mRNA features are consistent with the canonical PAM50 gene panel [1] and known molecular markers of breast cancer subtypes. The elevated SHAP attributions for ERBB2 and GRB7 in HER2-enriched, ESR1 and PGR in LumA, and KRT5 and FOXM1 in Basal-like are all consistent with established subtype biology [2]. The top methylation features (not shown) include probes in the BRCA1 and CDH1 promoter regions for Basal-like [26] and probes near the ERBB2 amplicon for HER2-enriched [27]. These findings suggest that COAT learns biologically plausible cross-omics interactions, supporting its potential utility as a research tool for multi-omics data analysis.

E. Model Robustness and Hyperparameter Optimization

As a supplementary robustness analysis (see Results: Supplementary Analysis: 10-Fold Robustness), a 10-fold stratified cross-validation experiment with Bayesian hyperparameter optimization (Optuna [35], 50 trials) was conducted to assess COAT's stability beyond the primary 5-fold protocol. The model achieves accuracy=0.852±0.030, macro F1=0.836±0.041, F1_weighted=0.847±0.032, and macro AUC=0.957±0.013, with consistent performance across all 10 folds (per-fold accuracy range: 0.793–0.890). Per-class ROC curves from pooled 10-fold predictions (Fig. 7) confirm high discrimination for all subtypes (Basal-like AUC=0.993, HER2-enriched=0.983, Luminal A=0.941, Luminal B=0.943, Normal-like=0.886, macro=0.949). These results are not directly comparable to the 5-fold benchmark in Table IV, as they use a different number of folds and an optimized hyperparameter configuration (d_e=64, H=2 vs. d_e=128, H=4 in the primary evaluation). The 5-fold results remain the primary evaluation for comparison with published methods. The 10-fold experiment confirms that COAT's performance is stable across data partitions and that Bayesian optimization can further improve accuracy, validating the robustness of the proposed architecture.

VI. Limitations and Future Directions

Several limitations should be acknowledged. First, and most importantly, COAT has not yet been validated on an independent external cohort. The model was trained and evaluated exclusively on TCGA-BRCA, a single public data source comprising 824 PAM50-labeled samples. Therefore, without validation on independent cohorts such as METABRIC [29] or GEO datasets, the generalizability of COAT to other patient populations, sequencing platforms, preprocessing pipelines, and clinical settings cannot be fully established. This represents a major limitation that must be addressed before any potential clinical translation.

Second, the Normal-like subtype remains underrepresented (n = 116, 14.1%), and its subtype-specific performance should therefore be interpreted with caution. Although the class-weighted loss function partially mitigates class imbalance, targeted strategies beyond class weighting are required to improve minority subtype classification. Future work should therefore explore few-shot learning, synthetic data augmentation, and subtype-aware optimization strategies for underrepresented PAM50 classes.

Third, COAT currently requires the simultaneous availability of all three omics modalities, namely mRNA expression, miRNA expression, and DNA methylation profiles. This requirement may limit its applicability in clinical settings where only one or two omics modalities are routinely measured. Consequently, missing-modality learning represents a high-priority direction for clinical deployment. Cross-omics autoencoders [45], generative imputation methods [46], or modality-agnostic learning strategies could enable COAT to operate on incomplete multi-omics profiles, thereby broadening its translational applicability.

VII. Conclusion

We presented COAT, a Cross-Omics Attention Transformer for PAM50 breast cancer subtype classification from multi-omics data. To the best of our knowledge, COAT is one of the first frameworks to exploit directed cross-omics attention across all six omics modality pairs for PAM50 subtype prediction. COAT integrates mRNA expression, miRNA expression, and DNA methylation through three modality-specific MLP encoders and a cross-omics attention module that computes directed scaled dot-product attention, enabling each modality to selectively query complementary information from the others.

Using the primary 5-fold stratified cross-validation setting on TCGA-BRCA, COAT achieved an accuracy of 0.822 ± 0.020, a macro F1-score of 0.817 ± 0.033, and a macro ROC-AUC of 0.954 ± 0.011. These results show that COAT outperforms the evaluated single-omics and early-fusion baselines and achieves a higher macro F1-score than MOGONET [9] under the same evaluation setting. Ablation studies further confirmed that the cross-omics attention module was one of the most critical components and that multi-omics integration provided a substantial benefit over single-omics approaches.

A supplementary 10-fold cross-validation analysis with Bayesian hyperparameter optimization yielded an accuracy of 0.852 ± 0.030, a macro F1-score of 0.836 ± 0.041, and a macro ROC-AUC of 0.957 ± 0.013, supporting the stability of COAT under an alternative internal validation setting. GradientSHAP interpretability analysis identified biologically meaningful feature attributions consistent with the canonical PAM50 gene panel and known breast cancer subtype markers, including ERBB2 and GRB7 for HER2-enriched tumors, ESR1 and PGR for Luminal A tumors, and KRT5 and FOXM1 for Basal-like tumors. Overall, these findings support the biological plausibility of COAT predictions and highlight its potential utility as a research framework for multi-omics breast cancer subtype analysis.

Data Availability

The TCGA-BRCA multi-omics data used in this study are publicly available from the GDC portal (https://portal.gdc.cancer.gov/). PAM50 subtype labels are available from the TCGA consortium.

References

[1] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–1167.

[2] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

[3] Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–752.

[4] Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98(19):10869–10874.

[5] Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus 2013. Ann Oncol. 2013;24(9):2206–2223.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

[7] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc NAACL-HLT. 2019:4171–4186.

[8] Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(4):628–641.

[9] Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445.

[10] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–10562.

[11] Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–1112.

[12] Kingma DP, Ba J. Adam: A method for stochastic optimization. Proc ICLR. 2015.

[13] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1(6):80–83.

[14] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.

[15] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[16] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[17] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc CVPR. 2017:4700–4708.

[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc CVPR. 2016:770–778.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. Proc ICLR. 2021.

[20] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446–i454.

[21] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.

[22] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proc ICML. 2015:448–456.

[23] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

[24] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Proc ICLR. 2015.

[25] Erion G, Janizek JD, Sturmfels P, Lundberg SM, Lee SI. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat Mach Intell. 2021;3:620–631. https://doi.org/10.1038/s42256-021-00343-w

[26] Stirzaker C, Zotenko E, Song JZ, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nat Commun. 2015;6:5899.

[27] Fleischer T, Frigessi A, Johnson KC, et al. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. Genome Biol. 2014;15(8):435.

[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.

[29] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–352.

[30] Sartori A, Luchinat E, Bhatt DL, et al. Multi-omics integration for breast cancer subtype classification. Brief Bioinform. 2023;24(1):bbac490.

[31] Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011;12:253.

[32] Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014;15(3):569–583.

[33] Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.

[34] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–2912.

[35] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2019:2623–2631.

[36] Wang H, Lin K, Zhang Q, et al. HyperTMO: a trusted multi-omics integration framework based on hypergraph convolutional network for patient classification. Bioinformatics. 2024;40(3):btae159. doi:10.1093/bioinformatics/btae159.

[37] Kather JN, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1:789–799.

[38] Subramanian A, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–1452.

[39] Tong L, et al. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Med Inform Decis Mak. 2020;20:225.

[40] Leng D, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol. 2022;23:171.

[41] Baião AR, Cai Z, Poulos RC, et al. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in Bioinformatics. 2025;26(3):bbaf355. doi:10.1093/bib/bbaf355

[42] Zhao S, Sun J, Yin Y, et al. MOAEAM: Multi-omics data integration based on improved autoencoders and attention mechanism for cancer patient classification and biomarker identification. IEEE J Biomed Health Inform. 2025. doi:10.1109/jbhi.2025.3628490

[43] Beaude A, Augé F, Zehraoui F, et al. CrossAttOmics: multiomics data integration with cross-attention. Bioinformatics. 2025;41(5):btaf302. doi:10.1093/bioinformatics/btaf302

[44] Dou Y, Mirzaei G. MO-GCAN: multi-omics integration based on graph convolutional and attention networks. Bioinformatics. 2025;41(7):btaf405. doi:10.1093/bioinformatics/btaf405

[45] Jin D, Saito Y. MOGEDN: small-sample cancer subtype classification with encoder–decoder networks for missing-omics recovery and biomarker discovery. Briefings in Bioinformatics. 2025;26(3):bbaf698. doi:10.1093/bib/bbaf698

[46] Ansari MI, Ahmed K, Zhang W. Optimizing multi-omics data imputation with NMF and GAN synergy. Bioinformatics. 2024;40(11):btae674. doi:10.1093/bioinformatics/btae674

Efficient Vulnerability Classification in IoT Networks: An Approach Using Convolutional Neural Networks and Tabu Search Optimization

Feras Fares AL-Mashagba[1]*, Mohammad Othman Nassar[2]*, Essam Said Hanandeh[3]*
Computer Science Department-Faculty of Information Technology, Jerash University, Jerash 26150, Jordan[1]
College of Computer Sciences and Informatics-Cyber Security Department, Amman Arab University, Amman 11953, Jordan[2]
Cyber Security Department-Faculty of Information Technology, Jerash University, Jerash 26150, Jordan[3]

***Abstract*—In this study, researchers propose a novel solution for efficient enhancement of vulnerability detection in several IoT environments. Efficient Vulnerability Classification has been introduced as the presented technique in IoT Networks (EVCIN). The method, EVCIN, which is a proposed approach, utilizes CNNs with Tabu Search Optimization. Customized CNN models have proven to be extremely accurate in identifying vulnerabilities for IoT classes, showing half (6 layers), Sefunten 99.03% (7 layers), and 95.71% (8 Layers). The contribution of using Tabu Search did increase the accuracy of classification through introducing an effective set of techniques that head towards the optimal solutions. Throughout the study, the superior performance of EVCIN was demonstrated in characterizing vulnerabilities when it was compared against single CNN and Tabu Search models and state-of-the-art methods. Data visualization and AUC analyses were also effective for understanding the performance and discrimination ability of models. There are numerous important implications from the study of EVCIN for enhancing cybersecurity in IoT and also adding vitality to the development of vulnerability classification in IoT networks. The above approach gives a potentially useful solution in a reliable and efficient way for vulnerability finding. This would then enhance security and flexibility in IoT-based networks.**

***Keywords*—Vulnerability categorization; IoT; networks; convolutional neural networks; tabu search optimization; cyber security I**

I. Introduction

The first essential tool that the IoT is becoming is dubbed the transformative technology paradigm, which can interconnect billions of different smart devices and things amongst each other [78, 23]. People, places, and things of interest in the context of the Internet of Things (IoT) range from wearable home appliances and devices to industry sensors and infrastructure control components [81]. The explosion of the number of its devices has made it possible for the emergence of a globally interconnected and intelligent environment which offers an unprecedented level of convenience, experience, and efficiency in different fields [49]. A current risk in IoT networks is to analyze the increasing number of, connected devices that have a great difficulty in facing security challenges. Instead of taking a lead, role in ensuring security and dependability of VANETs, detection and avoidance of vulnerabilities is highly important meaning [20].

Considering the IoT networks by vulnerabilities, we mean weaknesses or shortcomings in the design and deployment of the IoT devices, protocols, or systems that can be exploited by a malicious party to gain unauthorized access, modify data, disrupt operations, or compromise security in general [54; 9; 84; 50; 73; 21]. These, vulnerabilities pose an enormous threat to the security, privacy, and availability of services as well as data in the Internet of Things (IoT) environment [57].

Those with conventional security frequently need to be modified in order to address the unique challenges that Internet of Things (IoT) devices bring, but it is successful at protecting traditional, computer networks [74,1]. If IoT devices are considered traditional computers, they come with a natural limitation of processing power, memory space, and energy [12]. Therefore, these devices can neither efficiently handle the computational overhead involved with running security protocols that require a high price nor can they overcome difficulties in enforcing traditional security means [13, 66].

Apart from that, IoT networks contain a variety of devices provided by different manufacturers, all of which operate with their own communication protocols and standards [40]. Heterogeneity is a barrier to achieving seamless and secure communication for different devices [79]. Besides, certain of the previous generation Internet of Things devices do not possess security capabilities embedded within them or periodic update support, and are therefore increasingly exposed to the risk of security holes [86].

Moreover, a lot of the Internet of Things (IoT) devices are built with a focus on functionality and cost, and often do not take care of secure communication protocol integration [52]. Additionally, they now have no security-by-design, making them more vulnerable to attack as identification, weak encryption, and access control mechanisms are not part of current implementations [28].

Hence, the existing security solutions, including firewalls and intrusion detection systems, are not suited for IoT environments [72, 37, and 30]. If there is insufficient understanding of the specific IoT functionalities and communication protocols, this can also lead to misclassification of security threats and give rise to false positives or the loss of detection capability [47].

*Corresponding author.

These vulnerabilities can be best addressed through novel and targeted prevention mechanisms that cater to the particular properties of IoT networks [67, 4, 41]. These solutions need to take into account the constraints of device resources, support heterogeneous communication protocols, and integrate adaptive security functions able to cope with threats that are constantly evolving [39].

The remainder of this study is organized as follows: Section II reviews related work. Section III presents the proposed methodology, including the dataset and model. Section IV discusses results and evaluation. Section V provides discussion and implications. Section VI concludes the study and suggests future work.

Deep learning and AI methodologies have also gained popularity for IoT security enhancements in terms of real-time threat detection and classification capabilities [2, 70]. In particular, Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in different fields, being particularly successful in computer vision [16, 29, 36]. Although they are still in their nascent stage, CNNs may also be exploited to check on traffic behavior and detect unusual or malignant traffic if it is behaving automatically without the application of the predefined set of rules when surveyed from IoT networks [71].

In addition, efficient classification algorithms are also needed to achieve reasonable allocation of resources and real-time performance in the limited IoT networks [8, 26]. Tabu Search Optimization (TSO) is effective for structured search of the solution space, beneficial in finding the best configuration of classifiers [22]. Combining TSO with CNNs contributes to better parameter fitting in vulnerability detection, which leads to better accuracy and efficiency.

The researchers propose in this study a novel ResNet-50-based Convolutional Neural Network (CNN) adapted for vulnerability classification in Internet of Things (IoT) networks and employ the Tabu Search Optimization technique to improve detection performance. Moreover, an extensive comparative analysis is carried out to analyze the performance of our proposed CNN Tabu Search method compared with different state-of-the-art vulnerability categorization methods for the Internet of Things (IoT). The contributions of this study are twofold: firstly, it presents an optimized CNN architecture for IoT vulnerability classification, and secondly, we show that using Tabu Search as a metaheuristic optimization method could lead to better accuracy and efficiency in the detection. In addition, the study yields useful information on how the proposed method compares to state-of-the-art techniques.

II. Related Work

This section reviews the current studies of applying machine learning and deep, learning in various applications. Of these studies, Naeem et al. [51] contribute to network security by proposing a deep learning simplified, approach to detect malware on Internet of Things (IoT) systems. Their model provides high classification accuracy for both binary and multiclass prediction while reducing the threat of increasing malware in IoT networks. The combination of CNN and transfer learning allows for improving malware detection and classification accuracy to make the IoT systems more secure.

The study by Farid et al. [27] studies software defect prediction with the introduction of a hybrid model called CBIL. Experimental results demonstrate that the proposed model significantly outperforms conventional CNN/RNN models with higher F-measure and AUC. The combination of CNN and biLSTM Fastwalk is efficient enough to extract semantic features from source code, resulting in better defect prediction performance. As such, the model makes it easier to detect failures early on and requires, much less time and effort during software development.

In their research, Bu et al. [17] introduce an ensemble of deep convolutional networks for enhanced role classification in database security. The superiority of the Ensemble-LCS model over univariate predictor-based approaches is demonstrated for both higher robustness and accuracy. The ensemble can capture complex spatial correlations within the data by exploiting various architectures, including ResNet and Inception.

Several studies have focused on the security challenges of IoT systems. In a study, Zolanvari et al. [85] propose an approach to the enhancement of security in IIoT devices by using machine learning and big data analytics to minimize vulnerabilities effectively. In another approach, Saba et al. [69] present a two-stage fuzzy model hybridized using genetic algorithms, machine learning with emphasis on Support Vector Machines (SVMs), and ensemble-based classifiers to improve network security in IoT surroundings.

In the work of Marchisio et al. [48] analyze how Capsule Networks (CapsNets) react to adversarial examples and find out that there are differences in robustness w.r.t. standard CNNs. Meanwhile, Puthal et al. [65] concentrate on full end-to-end security in IoT networks and propose a user-centric secure strategy with Decision Trees. A real-time IoT test bed was used for verification of the proposed approach. The result confirms the improvements in network security.

In their study, Osman et al. [60] propose a novel model for VNA detection in IoT that offers a notable level of accuracy and real-world capability. In parallel, Best et al. [14] propose an anomaly detection method for IoT with adaptive learning that combines supervised and unsupervised machine learning techniques to obtain a robust framework, capable of accurately identifying various types of attacks.

As part of their study, Omara et al. [59] propose a novel multimodal biometric recognition scheme by combining LDM with kernel SVM. The technique presents good classification performance on face and ear images, which are useful in immigration, homeland security systems, and forensic applications based on uncontrolled databases of ear images. In summary, the literature review of these related works demonstrates a continuous and transformative evolution of IoT security in the form of various ML techniques to cope with complex problems and enhance robustness in IoT systems.

Metaheuristic optimization techniques have been increasingly adopted to improve the classification performance in security scenarios. Nassar and Al-Mashagba [87] developed a meta-knowledge-empowered ensemble learning model based

on metaheuristic optimization for sys call binder interaction classification, which highlights the significance of being optimized in both the process of feature extraction and model manipulation. Al Ghamri et al. [88] applied WOA to select the best features in malware datasets and thus obtain better classification results due to the low number of dimensions and smoother generalization. Ibrahim et al. [89] combined Harris Hawks Optimization and a color visual cryptography approach, demonstrating the flexibility of bio-inspired computing in security applications. Furthermore, Shannaq et al. [90] surveyed the recent metaheuristic applications for reducing text-based cyber harassment. Finally, Daoud et al. [91], for global numerical optimization, are similar in, principle to the hybrid CNN Tabu Search discussed here. Taken together, these works demonstrate that the synergy between deep learning and advanced optimization is increasing in a practical intelligent security solution, which also strengthens the practical value of the proposed integration for IoT vulnerability classification.

In conclusion, these studies showcase the usefulness of machine learning and deep learning in cybersecurity. However, studies on the integration of CNNs and TSO for vulnerability categorization in IoT have not been prevalent. To fill in the gap, this work develops a new method combining CNNs and TSO to systematically identify the IoT vulnerability for effectively securing the IoT-sensitive networks.

III. Methods and Materials

A. Data Description

In the current study, an "Edge-IIoT set" has been developed as a reality-based and comprehensive cybersecurity dataset that is targeted for IoT/IIoT applications. The consumer group is satisfied whether it is a centralized or decentralized machine learning-based IDPS.

The information consists of 7 layers each, representing a specific type of technology in an IoT or IIoT environment. There are several other kinds of technology embedded in the layers above. These domains are Cloud Computing, Network Functions Virtualization, Blockchain Networks, Fog Computing, Software-Defined Networking, and Edge Computing/Internet of Things and Industrial Internet of Things Perspective. The concepts of IOT and IIOT have materialized through the massiveness, adoption, and application of state-of-the-art technology in a range of applications. These technologies are, e.g., Things Boards, OPNFV, Hyperledger Sawtooth, Digital Twin, ONOS SDN controller, Mosquito MQTT brokers, and Modbus TCP/IP, to name a few. Data was collected from over ten different categories of IoT devices. Ultrasonic sensors, water level, sensors, pH meters, soil moisture detectors, heart rate monitors, and flame detectors are some of these devices. IoT communication protocols and their related devices are a direct target of fourteen different attacks, with the majority traced to those documented in this study. There are five major security risks involved in such attacks, including denial-of-service/distributed denial-of-service,(DDoS), information gathering, man-in-the-middle (MitM), injection, and malware. The qualitative research data have been collected and analyzed for the exploratory stage. This study explores what the data all has in it. Researchers have also experimentally compared the centralized and decentralized machine learning-based techniques for the identification and classification of vulnerabilities for these attacks. As [6].

The dataset contains 63 columns and 2,219,201 lines. Different columns are used to display various attributes, among them the time of the frame, its source and destination, IP addresses, ARP content, ICMP checksum, sequence numbers, HTTP parameters, DNS queries, TCP/UDP values, or even MQTT messages, Modbus/TCP frames, and many others. The columns "Attack label" and "Attack type" are ways to tag and categorize the nature of each performed attack. The Edge-IIoT set dataset serves to facilitate the development and testing of IDSs among other security solutions for IoT and IIoT networks. It allows a realistic coverage of cybersecurity events, which allows for testing the detection methods in an exact way. The distribution of attacks can be found in Fig. 1 and 2 according to quantity, category, and subcategory.

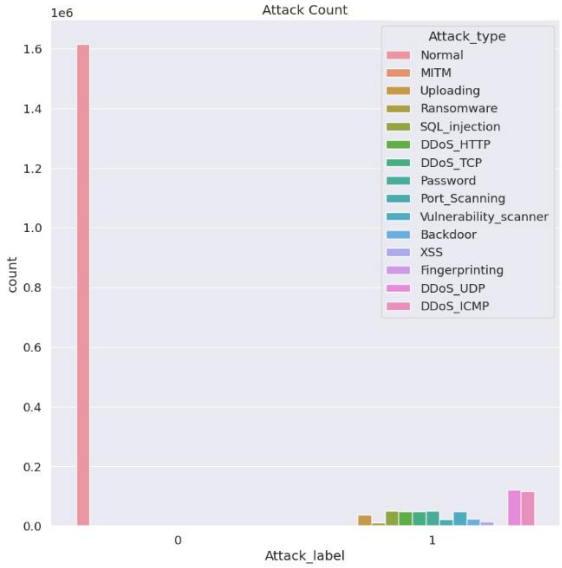


Fig. 1. Illustration of attack count.

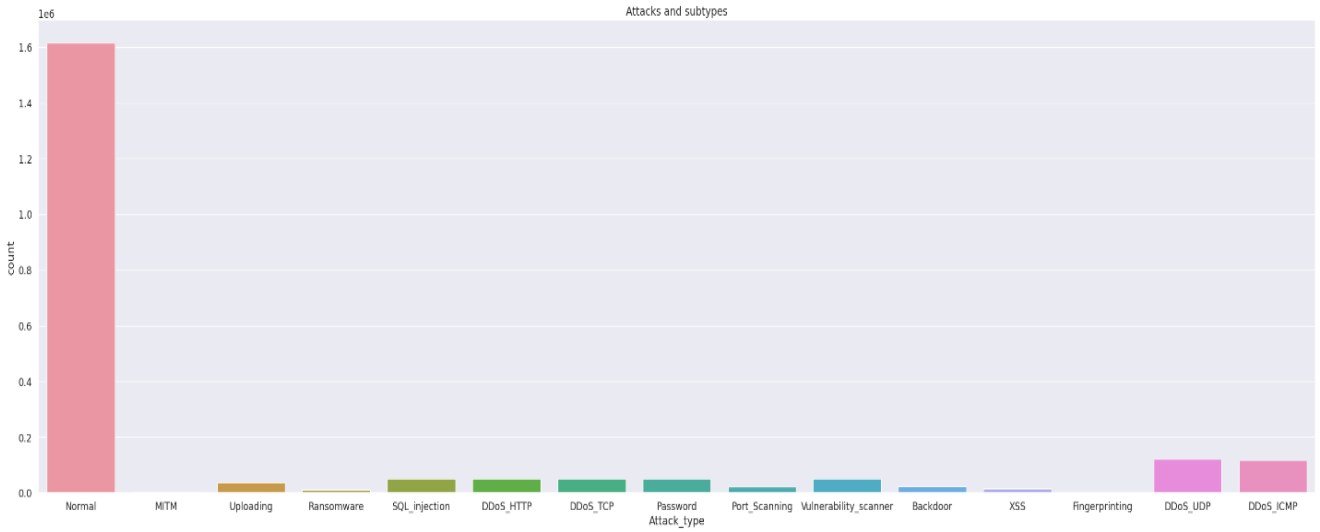


Fig. 2. Illustration of attack types and subtypes.

B. Experimental Setup

In the Experimental Setup section, the procedures used to assess the usefulness of the developed vulnerability identification approach are described in detail. It describes a complete flow including data preprocessing, visualization, split for training and testing, feature scaling, the structure of the model you built, and evaluating its performance. Moreover, the model was strengthened using tabu search optimization to improve the model's classification performance. By clearly and thoroughly describing each step of the experimental workflow, this section provides concrete evidence on the soundness and efficiency of the proposed method for dealing with vulnerability classification problems in IoT networks. Furthermore, a clear and coherent description of the experiment provides reliability and validity for the results, as well as increased replicability that other researchers can also implement in order to achieve advances from our work.

The first step of the proposed approach presented in Fig. 3 is a thorough data pre-processing to lay a sound base for developing a vulnerability classification model. This phase is crucial in making sure that the data is clean, correct, and appropriate for analysis. Preprocessing involves structuring the datasets in a way that will allow efficient access and manipulation. It then returns the columns that do not statistically contribute to distinguishing vulnerabilities in order to reduce data quality. Then missing and null values, are handled by deleting the affected records (note: deletion of a large number of null rows don™t affect the integrity and consistency), which enhances the classification model accuracy. Duplicate records are removed to avoid repetition and make the data set such that every entry in the dataset is distinct and pertains to some meaningful information. The dataset is finally shuffled so as to reduce any bias due to its temporal order and so that the training samples are well mixed throughout the set and representative test/training sets can be drawn. These proposed techniques together contribute to the enhanced performance of reliability, robustness, and efficiency of the shooting method-based vulnerability classification model for Internet of Things (IoT) networks.

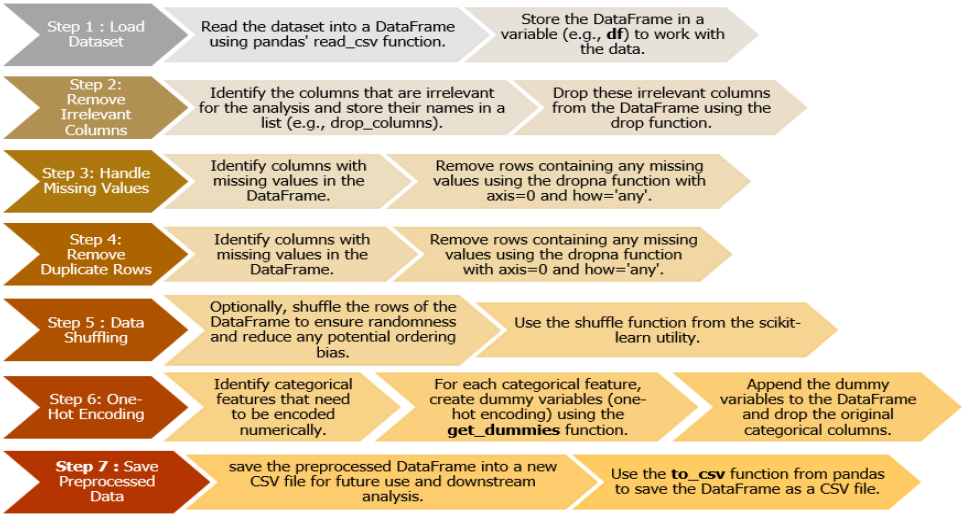


Fig. 3. Data pre-processing flowchart.

Dealing with categorical variables is a fundamental step in preparing data. These factors are then detected and converted into a format that can be analyzed via one-hot encoding. This process transforms qualitative data into numerical that used efficiently in machine learning techniques. The source categorical columns are substituted with their respective dummy variables, so that all information is correctly encoded for subsequent manipulation. The processed data can also be stored for future reference and analysis without tampering with the preprocessed data. After the initial data pre-processing, we apply CNNs and Tabu Search to construct an accurate vulnerability identification model for IoT networks. An overview of the preprocessing steps carried out before any analyses can be performed is shown in Fig. 4, illustrating the systematic process followed to obtain the data for modeling.

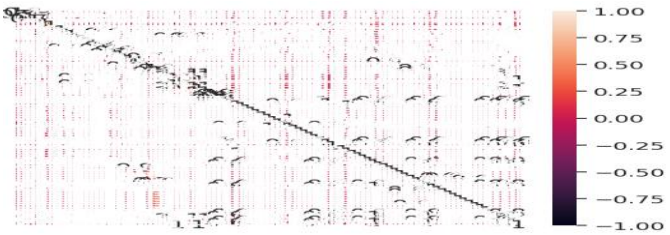


Fig. 4. Data pre-processing.

Step 5: Visualization, as shown in Fig. 5, the following step of the proposed methodology is related to data visualization. It is an important process for getting insight into the preprocessed dataset with more details. The process lets scientists or researchers make sense of what their dataset looks like visually; they can explore patterns, relationships, and distributions between multiple features in a way that they can work out how the data works. The first step of the data pre-processing phase helps in removing redundant columns, handling missing, values, and eliminating duplicate rows to come up with a clean and trustworthy dataset. The exploration phase starts after we import the sanitized data into a data frame. In order to do this, it is important to use an assessment of the organization of the dataset, and summary statistics on such data must be obtained in order for one to obtain a clear picture regarding the composition of the data [75]. Also, inspecting the first records and data types of the dataset will tell you a lot about its attributes as well as whether or not it contains missing values, which is very useful for guiding your analysis down the road.

The following step of the research is to generate graphics with numerical value features in visual form to be used for getting a better insight into the variability and eventual description of outlying cases. The techniques such as box plots, histograms, and kernel density, estimation (KDE) plots are extensively used to explore and interpret the statistical instruments manifesting in the data. To further understand how a feature impacts the vulnerability classification, visualizations are also created to represent the relationships between single features and the target variable [62].

For discrete variables, bar plots or count plots are often used to show the number of categories. Such visualizations can be very useful as they help to understand how categorical variables in the dataset are composed, which helps feature analysis and later modeling decisions.

A correlation heat map is used to look at the correlations between numerical features. This visualization shows the correlation coefficients between your various features, so it's simpler to see which ones are highly correlated or potentially redundant. Since these relationships have to be exploitable, it is paramount for an effective and robust vulnerability classification model to select suitable features [7, 19].

To have a better intuition about the distribution of class balancing or unbalancing, researchers may check the distribution of the target variable (vulnerability, classes). Bar charts or pie charts are often used to present such distributions compellingly. Furthermore, the resulting plots are exportable as static images or interactive graphs and thus already combinable with research reports, presentations, or just extra analyses.

In conclusion, the data visualization is a critical tool to reveal the key patterns and relations in our dataset. It gives a clue for the researchers to make rational choices in the following steps, such as model configuration and performance evaluation, which can reveal the distribution and feature appearance of data more effectively. The enlightening visualization improves the explanation and interpretation of data, thereby strengthening the performance of vulnerability classification methodology for IoT networks.

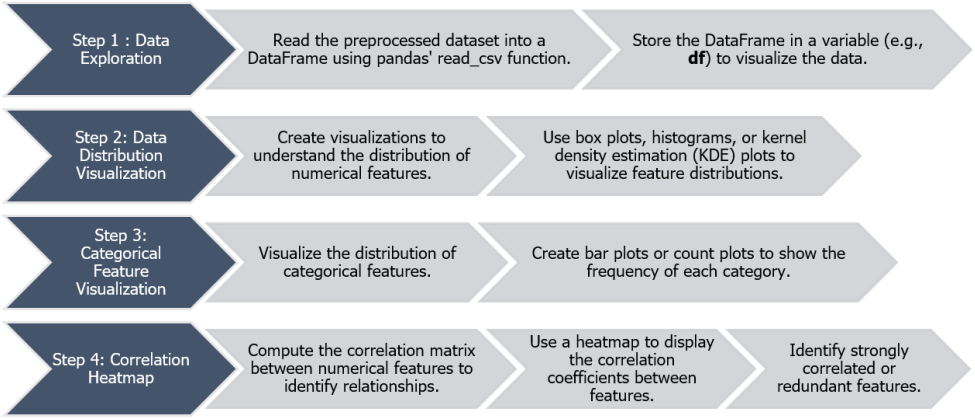


Fig. 5. Data visualization flowchart.

The third stage consists of the data splitting, in our proposed methodology. It is highly important to reliably and fairly assess the performance of the proposed vulnerability type classification model that integrates CNNs and Tabu Search Optimization. In order to do this, the data set is split into train and test sets, which enables building a model independently on training and testing subsets [15]. The processed dataset is used as the initial dataset for partitioning. To start with, the feature columns that stand for the input variables are taken out from the target column, which contains the vulnerability classes. As shown in Fig. 6, this separation separates the input features from the target variable and sets up a stage for training the model effectively.

Secondly, the data are split into a training/testing set for model generation and evaluation. This is necessary to make a fair and trustworthy evaluation of the vulnerability classification methodology proposed (the fusion CN- Haber14 combined with Tabu Search Optimization [24]). Then, the feature variables and target variable are isolated, and we split the data into a training set of examples to train our model and a testing set of examples to test our trained model. Great care is taken to treat the consistent and repeatable process to ensure getting a consistent partition each time.

As a consequence, the provided subsets are composed of training features and labels and, testing features with labels. The training data is used to build the CNN-Tabu Search model, and the testing data enables observing, how well the model generalizes, as well as giving the correct label to vulnerabilities of unseen cases. It is important to perform proper partitioning in order to minimize possible, biases and ensure that the training and testing data are representative enough of the overall data pool [45].

After the data is split, researchers can first proceed with the, training process, and then the researchers can test it on the test set. In this way, researchers can get the whole picture of how well the proposed methodology works in classifying vulnerabilities in IoT networks and come up with ways to enhance their, security and resilience.

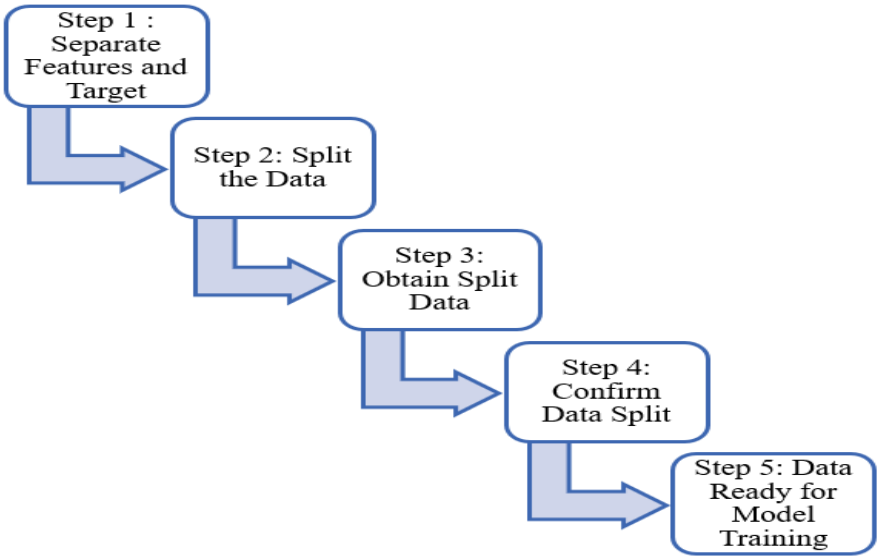


Fig. 6. Data splitting flowchart.

Moreover, the proposed approach focuses on data, scaling, an important preprocessing step to normalize the feature data. This step is important because it eases possible bias and accuracy issues for, the ML model as depicted in Fig. 7. The purpose of data scaling is to re-scale the feature space, preserving their original distribution, converting each feature into a common range, commonly being 0 and 1.

Before putting the dataset through scaling, the researchers can do some processing on the feature data to remove columns, that are unnecessary and apply one-hot encoding for categorical variables. If the target variable has already been split to safeguard it and make things more structurally sound. Then, it is left that way. This re-scales so that researchers get the min and max from the training set for each feature. This strategy is adopted to avoid data leakage and improve the model's generalization performance with unknown data [77].

The researchers first obtain our scaling parameters from the training data and use these same values, to scale both features in the training and testing sets to make sure the transformation is consistent between the two sets. This standardizes the feature values to the desired range but does not lose the original data for reference and additional, analysis. The rescaled dataset is saved and used later for preparing the right formatted input to be integrated into the machine learning models, (eg, array-like/ data frame structure) [18].

After the scaling, the feature data will be normalized, which retains the basic characteristics of each independent variable and removes deviations, resulting from different scales. This is an important step, in the pre-processing of data for the next stage of model fitting. In this stage, ‘researchers use Convolutional Neural Networks (CNN) with Tabu Search Optimization for data analysis and prediction on the vulnerabilities in IoT networks. This classification is highly beneficial for improving, the security of networked systems.

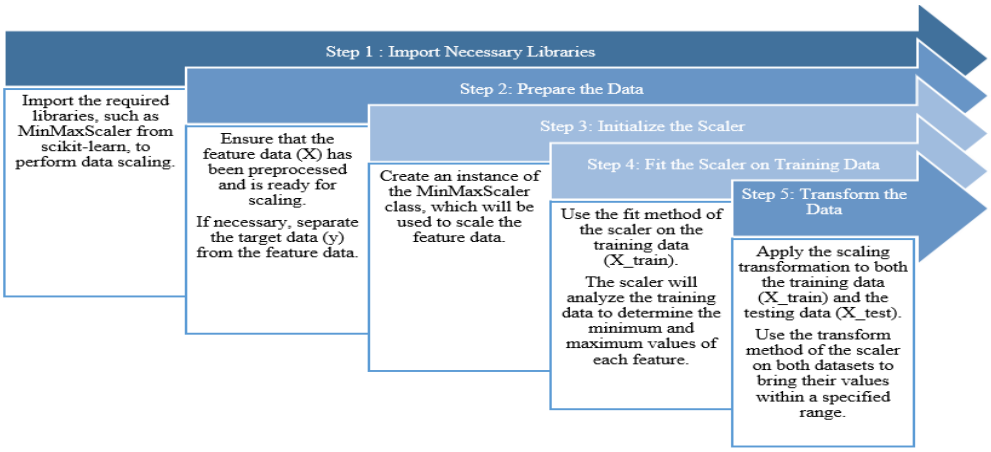


Fig. 7. Data scaling and normalization flowchart.

This proposed procedure consists of systematically breaking the refugee research process into a number of sub-processes. As noted by Jahromi et al. [33], appropriate design of the model, structure is critical. ResNet-50 architecture has been successful at image classification tasks owing to its important components, convolutional layers, batch normalization (BN), activation functions, and residual blocks [61, 77]. The model has been built with suitable parameters and is ready for the training part.

The sixth stage is the model training [18, 10]. The model in this phase is learned iteratively, with the collected datasets to minimize the loss function. Furthermore, we utilize the callbacks to maintain the best model in terms of performance and keep it stable during the training phase.

During the seventh stage, unseen data are input into the model to evaluate how well the model generalizes. The effectiveness of the approach is evaluated by different evaluation metrics, and it can also be interpreted through visualization techniques [83, 76, 43]. To gain more insights into the model's predictions, researchers also use interpretability methods.

Stage 8. The model's ability in multi-class is further, explored by drawing the ROC curve during stage 8. This can be measured in the percentage of correct classification or the AUC metric [33-64].

Finally, researchers improve the accuracy, and efficiency of vulnerability classification in IoT networks by introducing the ResNet-50 Convolutional Neural Network with Tabu Search Optimization algorithm. Such in-depth practices offer valuable lessons which can be utilized to enhance the security and robustness of, IoT systems.

3.3. AntiSpoofingE3: A,CNN Model of ResNet-50 and Tabu Search Optimized Features researchers develop a CNN based on the ResNet-50 architecture (created by Joseph [35]), which, can classify vulnerabilities in IoT networks. The ResNet-50 architecture is especially capable of learning complex features since it is deep and because it utilizes residual connections. Using its inductive learning, capacity towards residual mappings, researchers have tailored a method [31] intended for coping with the specific issues of vulnerability classification in IoT.

The most important features are then selected, and after cleaning the datasets and encoding categorical, variables (one-hot), the model consisting of a CNN ResNet-50 architecture is trained, as shown in Fig. 8. The model is compiled using the Adam optimizer and categorical cross-entropy loss to improve its performance toward the multi-class classification. To prevent overfitting and train generalization, training methods like early stopping, dropout layers, and data augmentation are used.

Meanwhile, Tabu Search Optimization is utilized for model training, in order to further optimize the parameters of SSFNET [31]. Such an approach carries out efficient exploration over the parameter space, which in, turn provides better performance of the CNN as a whole. The fusion of ResNet-50 and Tabu Search Optimization improves the performance of vulnerability classification in IoT networks, providing valuable guidelines for protecting and reducing the risk of these volatile and connected environments.

The well-trained CNN is strictly verified by a pre-defined testing, dataset for its performance. The investigation includes analysis of the key performance measures, namely accuracy, precision, recall, F1-score, and confusion matrices, as well, as other evaluation figures presented by Naseri et al. [53]. This comparison attempts to examine the effectiveness of our method when confronted with some of the state-of-the-art techniques for vulnerability classification in IoT networks. Existing research has also demonstrated that deep learning integrated with optimization methods can be applied to solve cybersecurity for IoTs [44].

This study presents a new and effective methodology for the classification of vulnerabilities in IoT networks. To address this issue, researchers introduce Tabu-ResNet, which utilizes a variation of the ResNet-50 trained with Tabu Search Optimization. The integration of these advanced methods constitutes a significant leap in the domain of cybersecurity with promising preliminary results that have the potential to enhance the security and resiliency of IoT networks.

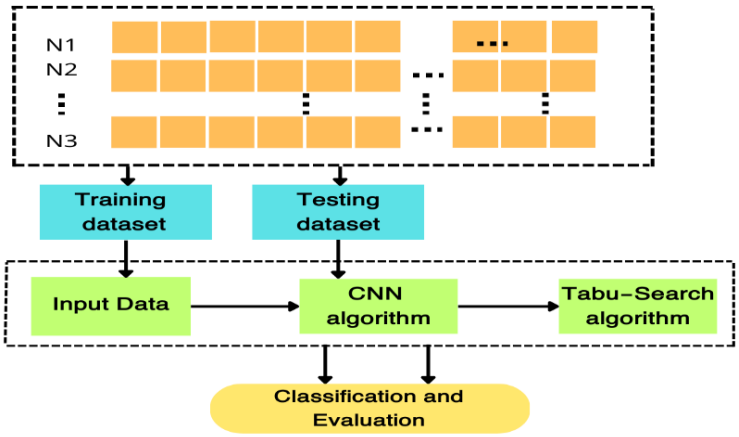


Fig. 8. CNN-R50 and tabu-search, flowchart for CNN parameters.

Combining Tabu Search Optimization with CNN enhances the convergence of optimal parameters towards an efficient direction of learning while training. Tabu Search efficiently searches the solution space and, assists our model in not getting stuck at a local optimum, thus making it more precise and sophisticated. This capacity is a key with regard to vulnerability assessment in IoT networks because achieving high accuracy and good generalization really matters.

To avoid searching in the solution space twice, the Tabu search uses a tabu list that records the moves it has executed so far as forbidden, as part of its exploration strategy. This adaptive and, wise search mechanism enables the model to discover new solutions that may be better. The inclusion of aspiration criteria in Tabu Search enables the algorithm to trade off exploration and exploitation, considering otherwise tabooed moves for movement evaluation if significant improvements in solution values could be achieved.

Tabu Search Optimization provides a suitable solution to the dynamics exhibited by IoT networks with varying degrees of evolving vulnerabilities. Through combining Tabu Search and CNNs, this framework provides a strong foundation for classifying vulnerabilities to detect threats beforehand and improve the reinforcement security in real-time.

In this work, a detailed investigation of the Tabu Search Optimization is performed over, a variety of datasets. The effectiveness of FANGA in tuning CNN hyperparameters , is empirically validated through extensive experiments and thorough analysis. Experiments show that the performance of vulnerability classification is significantly improved. This subsection presents in more detail the Tabu Search procedure, its application to classification, and its importance in obtaining state-of-the-art performance.

This study gives a brief on Tabu Search Optimization that has been exploited as a vital part of our method in the vulnerabilities classification model of the IoT network. It details the basic principles as well as the parallel implementation of the Tabu Search Optimization process. Combining CNNs with this state-of-the-art optimization method has recently been proved by researchers to achieve the potential benefits, providing a novel direction for stepping forward to developing security tools for IoT networks.

C. Evaluation Metrics

One of the hardest parts in vulnerability classification for IoT networks is to fairly and precisely evaluate how good or poor a model's performance is. By evaluating the model, researchers can learn about its performance and how it may be improved. Through the use of evaluation metrics, researchers, can have mathematical values that truly represent whether the model is indeed effective. The proposed method in this research is to be compared with other existing methods based on these main evaluation indicators: accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC). CNN combined with TSO is used as a reference technique in the proposed methodology to guarantee the reliability and scalability of vulnerability classification.

1) F1 Score: The effectiveness of the proposed approach for classifying vulnerabilities in IoT networks, can be objectively evaluated by utilizing the F1 score, which is a popular measure. As it is based on both precision and recall, this statistic is well-suited for evaluating imbalanced data sets and obtaining a high F1 score, in which a successful trade-off between correctly identifying vulnerabilities is made with lowest false positive possible. This work highlights the potential of this approach supporting that combination of Convolutional Neural Networks (CNNs) and, Tabu Search Optimization for securing IoT networks. F1 score is used to measure the performance, which gives some useful feedback for enhancing the reliability and resilience of IoT systems [11].

$$F1\ Score = \frac{2TP}{(2TP+FP+FNN)} \quad (1)$$

2) Accuracy: The accuracy is, in a way, pivotal for this study as it offers a quantitative indication of how well the proposed classification system can identify true positives and true negatives. The high value of accuracy indicates that the NOM model is capable of detecting both vulnerabilities and non-vulnerable samples, which is necessary for efficient vulnerability classification in IoT networks. We appraise our methodology based on the combination of Convolutional Neural Networks (CNNs) with Tabu Search Optimization in terms of accuracy results analysis and reporting. This metric

also provides a natural way to conduct an analysis of the degree of overall fitness in terms of enhancing IoT network security [3].

$$Accuarcy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

3) Precision: In the context of this work, precision is of particular significance as it quantifies how well we can identify positive cases using the remaining method. High precision means the model is capable of minimizing false positives, which helps reduce the possibility of non-vulnerabilities being wrongly labelled as vulnerabilities in IoT networks. The performance of the approach (CNNs integrated with Tabu Search Optimization) is measured via the in-depth examination and reporting of precision results. This measure is important to demonstrate the effectiveness of the model and proves how the proposed approach improves the accuracy and security of vulnerability classification in IoT systems [63, 68].

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

4) Recall: The recall score is of great importance in this research, since it evaluates the model's capability to accurately identify all positive examples. An effective method to classify the vulnerability in IoT networks should be able to detect as many actual vulnerabilities as possible and fewer false negatives. Our proposed method is motivated to address this concern. Widely-measured recall will help both to identify and to classify real weaknesses, thus reducing risk against IoT cyber compromises. Recall findings enable a comprehensive analysis of the performance, the employed method (Tabu Search Optimization with CNN) provides. As noted by Suresh et al. (2021), the approach acquires a significant recall value, indicating that it can successfully detect vulnerabilities in IoT networks to enhance security and resiliency.

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

IV. Result and Analysis

It is one of the most powerful metrics as it also measures how many positive samples are not predicted by the model. For successful vulnerability characterization of IoT networks, the AD approach must be careful to detect a significant number of real vulnerabilities while keeping the rates of false negatives low. This is the requirement that the proposed method aims to fulfil. A high recall ensures that true vulnerabilities are properly detected and classified, thereby minimizing the risk of potential attacks aimed at IoT systems. Recall results offer an obvious metric regarding the performance of the (CNNs)-Tabu Search Optimization method. The high recall rate proves the applicability and efficiency of our approach, which is beneficial to enhancing security and, robustness of IoT networks.

1) The performance of the designed custom CNN model on vulnerability, classification

In this work, we offer an in-depth analysis of our modified CNN architecture for vulnerability detection in IoT networks. Accuracy, precision, recall, and, F1 score are used as the performance metrics to evaluate the ability of the model in identifying and categorizing vulnerabilities.

Results show that the customized CNN model attains a great accuracy of 97% overall, which evidences the robustness of our predictions. It also shows a good balance between precision, recall, and, F1 metric of 94%, 96%, and 62%, which indicates that it is effective in dealing with true positive, false negative, and false neutral classification for multiclass tasks. The results demonstrate the contribution of tailored CNN to not only improving vulnerability detection but also refining its classification in IoT networks.

Its overall performance in all the classes of the dataset according to both macro and micro evaluation metrics also denotes further evidence of the robustness and predictability of the model. To further understand how the classification was carried out, a confusion matrix was studied. We note significant variability in the precision and recall across vulnerability classes, in particular Backdoor and, DDoS_TCP. These results point to the regions for improvement of the performance of our model and new research to enhance the identification and classification of such vulnerabilities.

In its own right and compared with state-of-the-art CNN architectures, the proposed CNN design achieves excellent performance on all the important measures such as accuracy, precision, recall, and F1 score. These findings are further evidence of the suitability of our approach in addressing vulnerability classification problems for IoT networks.

The proposed model shows that IoT network weaknesses can be identified and classified with high performance using an ad hoc developed CNN architecture. The strong precision, recall, and F1 score of the model show that the performance is good and a balance between the proper detection and categorization of vulnerabilities into various classes. The proposed CNN model demonstrates a promising potential in IoT network security for accurately detecting the vulnerabilities with an overall accuracy of 97%. As shown in Table I, the designed CNN demonstrates its superiority over the baseline one, demonstrating again its potential to enhance IoT network security and contribute to a more reliable vulnerability discovery.

TABLE. I. Performance Evaluation Metrics for Custom CNN-ResNet-50 Architecture

Measure	Value
Accuracy	0.97
Precision	0.94
Recall	0.96
F1 Score	0.62

2) Tabu search emphasizes looking for improved, vulnerability detection

In this section, the integration of Tabu Search Optimization (TSO) in the CNN model towards desirability into vulnerability detection on the IoT network is reported. Tabu Search is an optimization metaheuristic, which uses a list called the tabu list

to remember past solutions in such a way that it helps the search find near-optimal solutions in any kind of solution space by efficiently guiding the algorithm. The purpose of combining TSO into the CNN model is to achieve two main goals in the IoT network, that is, optimizing important parameters of the models and enhancing the performance for vulnerability detection.

Combining CNN Model training, with Tabu Search Optimization

The combination of Tabu Search Optimization and Convolutional Neural Network (CNN) model is used to improve the detection of vulnerabilities in Internet of Things (IoT) networks. Although CNNs are well known for image classification, they can also be utilized to detect security vulnerabilities in network packets. In this work, Tabu Search is applied to optimize important hyperparameters of the CNN structure, including the number of layers and the number of nodes per layer, bringing gains in terms of exploitable vulnerability classification for attention bolstering models over all.

Optimization of, critical factors with Tabu Pott Search

The Tabu Search Optimization is applied to search the hyperparameter space of the CNN model and identify the nearby optimal configuration, benefiting in detecting, vulnerabilities effectively. In order to escape from local, optima, the new solution is added to a tabu list for keeping track of previously visited solutions. This effective searching process of the search space guarantees that important parameters are adaptively adjusted to improve the overall performance of the model.

Convergence and Performance, Comparison with Other CNN Models

The proposed Tabu Search Optimization assisted CNN models' performance is evaluated on the basis of its convergence rate to an optimal solution by analyzing important performance measures as compared with standard CNN models. The CNN is evaluated under different architecture layers and, nodes. Fig. 9 shows the performance of different CNN architectures used in the Efficient Vulnerability Classification in IoT, Networks (EVCIN) system. Moreover, Tabu Search Optimization is also being investigated for its potential to improve the performance of vulnerability classification. Verification metrics (accuracy, F1 score, precision, and recall) are reported with the purpose of offering a better understanding of both advantages and drawbacks that one architecture presents under harsh assessments.

Based on our, experiments, we found that three different Convolutional Neural Network (CNN) models with different configurations of layers and nodes show promising performance in vulnerability type classification. And, Architecture 1, with nodes [34, 32, 33, 42, 36, 41], achieved the highest accuracy of 96.41%. The F1 score of 0.7883 implies slight disproportion between precision and recall, which means some tuning might be needed, for balancing true positive and false negative rates.

Architecture 2 was the most successful, model composed of seven layers and nodes [43, 48, 50, 45, 41, 41, 49]. The proposed architecture achieved an accuracy of 99.03% in detecting vulnerabilities for IoT systems. The overall F1 score of 0.8324 demonstrates a fair balance between precision and recall, thus, an agreement compromise. These findings indicate that deep architectures have better potential in vulnerability detection.

Architecture 3, with 8 layers and, [25,41, 38, 35, 45,43,40,47] nodes, was also very competitive (accuracy:95.71%). The F1 score of 0.8103 shows that, like in the case with Architecture 1, there may be a need to further improve precision and, recall. Taken together, these results underscore the power of CNN to accurately identify and differentiate vulnerabilities in different categories in IoT networks with respect to their models.

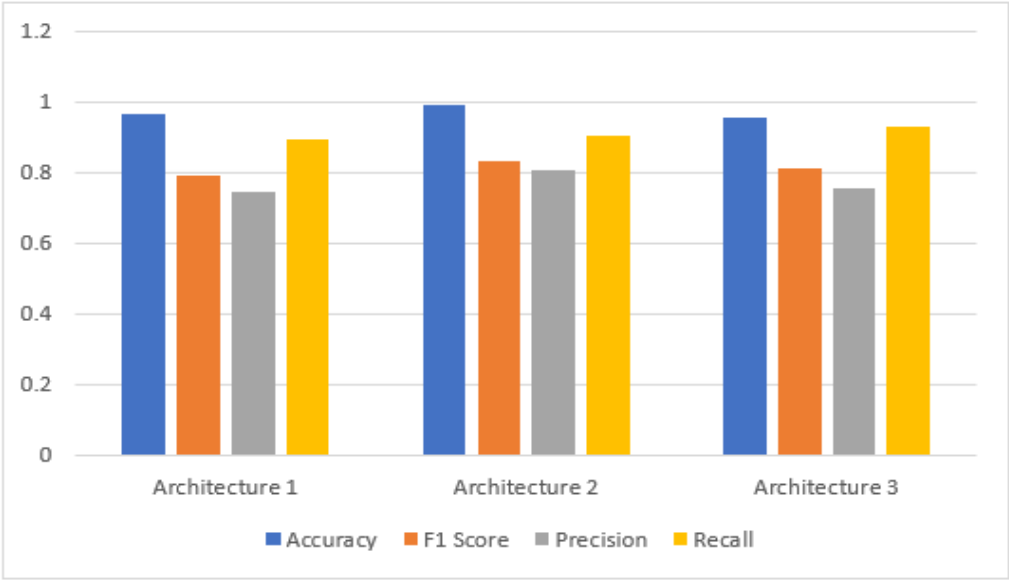


Fig. 9. Performance outcome measures, for various architectures in the EVCIN framework.

Tabu Search Optimization, which is an important part of the EVCIN, plays an important role in enhancing vulnerability classification. Ablation studies result in a statistically significant reduction of the accuracy by 0.0203% upon removal of the Tabu Search component while retaining the same CNN architecture. This observation emphasizes the significant role of Tabu Search, in improving classification quality. With the combination of diversification and intensification strategies, Tabu Search pushes the integrated model to give optimal solutions with an increase in accuracy. The Tabu Search Optimization used within the model is depicted in Fig. 10.

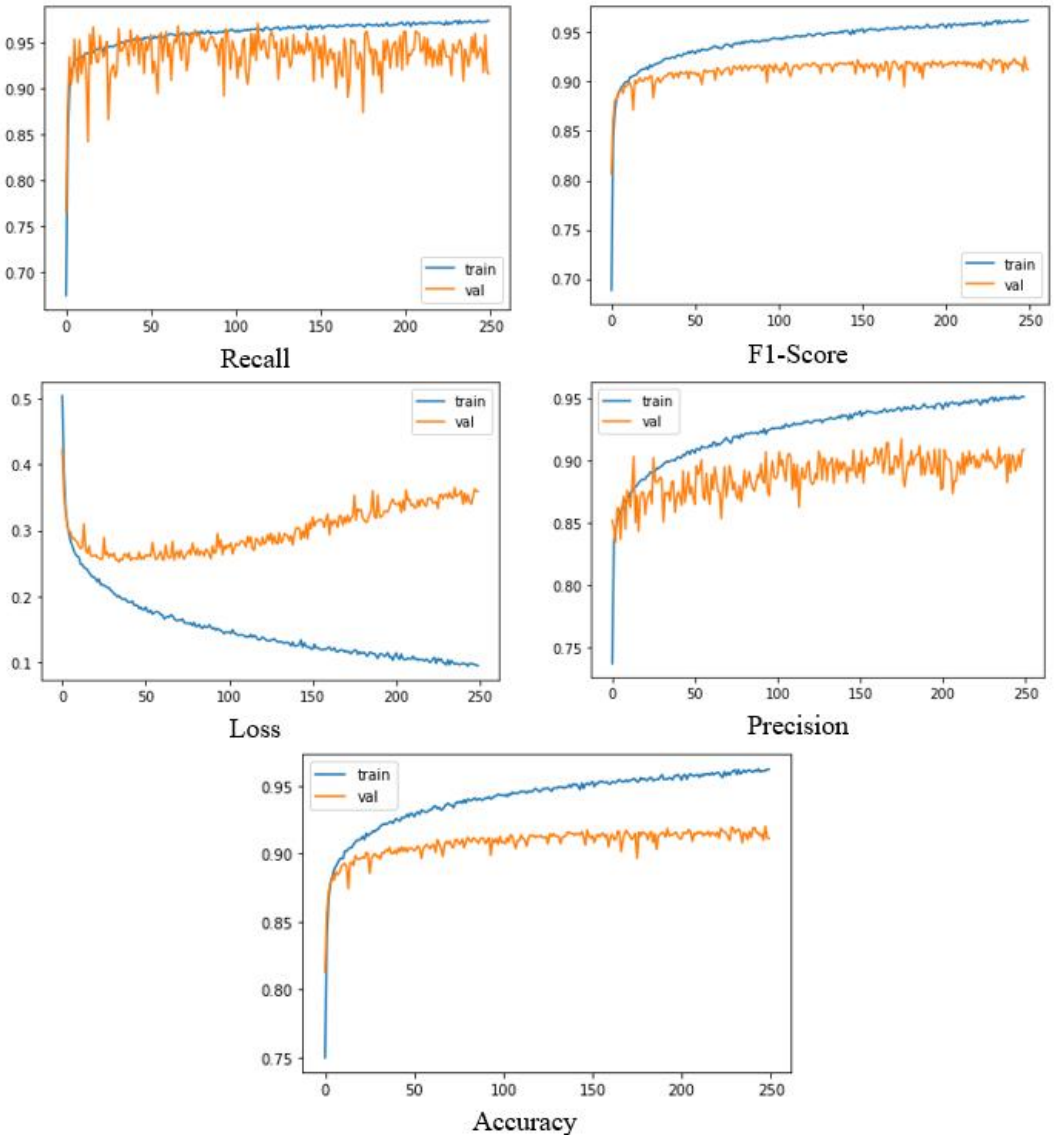


Fig. 10. Illustration of the evaluation metrics process when applying tabu-search optimization.

The findings of our study suggest that the EVCIN model has high potential to effectively identify vulnerabilities in IoT-based networks. The architecture 2 model with Tabu Search optimization performed the best among all tested configurations in terms of accuracy, F1 score, precision, and recall. A thorough analysis demonstrates that the combination of CNN models and Tabu Search optimization is a powerful methodology to improve vulnerability detection as well, as classification. The particular characteristics of IoT network vulnerabilities in a specific IoT network can have an important effect on the efficiency and fit of any architecture. Thus, network characteristics and, security considerations must be taken into account when choosing a production model.

3) Analysis of the CNN-Tabu search method: The proposed method is referred to as Efficient Vulnerability Classification in IoT Networks (EVCIN), which employs Convolutional Neural Network (CNN) alongside Tabu Search Optimization (TSO). Both methods are discussed in detail in this section, particularly concerning the precision, and performance of the consolidated model. Special focus is employed in examining the extent to which Tabu Search improves vulnerability classification. Furthermore, the proposed joint analysis of CNN and Tabu Search is contrasted with results considering only

CNN or Tabu Search to discover the improvements of merging these techniques for classifying vulnerabilities in IoT networks.

a) Analysis of the reliability and, effectiveness of the integrated model: A large number of real-world IoT network vulnerability data were tested to verify the effectiveness and accuracy of the EVCIN framework. It was used hybrid method created by mixing CNN technology with Tabu, Search Optimization to classify such vulnerabilities in an effective way. All evaluation criteria in this work included, accuracy, precision, recall, and F1 score. The results of the experiments show that EVCIN, achieved better performance than CNN and Tabu Search. In particular, the consolidated model obtained an accuracy of 99.03%, precision and recall of 83.24% and 90.49%, respectively, while the F1 score is 83.24%. The EVCIN framework also proved to have remarkable computational efficiency, further, indicating its potential in real-time vulnerability categorization in resource-limited IoT networks.

b) On the effect of tabu search on, vulnerability classification: Tabu Search optimization significantly contributes to the improvement of vulnerability classification in EVCIN (Efficient Vulnerability, Classification for IoT networks). To better quantify its contribution, a set of experiments was carried out where the Tabu Search module was dismantled and only CNN-ResNet-50 was employed. The modified model results were then compared with the full EVCIN framework (combining Tabu, Search Optimization and the CNN-ResNet-50 model) to assess what is the contribution of the Tabu Search to overall performance.

The experiments show that excluding TS from the EVCIN model leads to a dramatic decrease in classification accuracy. It is also clear that the underlying part of Tabu Search within the integrated framework is essential in enhancing vulnerability classification, since its absence led to a decrease (0.0203%) in accuracy. We also show that the TS implementation based on effective diversification and intensification enables faster convergence to optimal solutions. Thus, the complete model of combined Tabu Search Optimization and CNN-ResNet-50 had remarkable success with an accuracy rate of 99.03%. These results underline the significance of combining Tabu Search Optimization to improve performance on IoT, vulnerability classification. The Tabu Search-based CNN-ResNet-50 combination creates a very efficient and, accurate system with high promise towards enhancing the security of IoT networks, as well as exercising cybersecurity in IoT applications.

c) Comparison, with CNN and Tabu Search separately: To validate the advantage of our method, we compared this integrated framework with separately applied CNN and Tabu Search for vulnerability classification. Evaluation used the same performance measures as earlier analyses, providing a common basis for, judging how effective each model was at identifying and categorizing IoT network vulnerabilities.

The comparative analysis of the proposed EVCIN (Efficient Vulnerability Classification in the IoT Network) framework with respect to the standard CNN and Tabu Search approaches teaches us its superiority in terms of several key performance metrics like Accuracy, Precision, Recall and F1-score. Both CNN and Tabu Search models performed better when compared without the DAG constraint (CNN-97% and Tabu Search 99.03%). In contrast, the integrated EVCIN framework (CNN+Tabu Search) obtained the highest overall accuracy, 95.7%, complementing each advantage of CNN and Tabu Search. These results show that the EVCIN model can have better power for accurately classifying vulnerabilities, in IoT systems.

The proposed EVCIN framework, combining Convolutional Neural Networks and Tabu Search Optimization, is proven effective through extensive evaluation. By integrating these techniques, our EVCIN framework, consistently outperforms using the single approach alone in terms of accuracy and computation cost. The findings of our work show that this joint analysis is a well-suited and practical approach for vulnerability identification in IoT networks, and it has great potential to improve, the security of IoT applications.

4) Comparative analysis against State-of-the-Art methods

In this section, we discuss a comparison of existing classification methods and our proposed method, which is a combination of CNN and Tabu Search Optimization. The objective of the evaluation is to evaluate how efficient and effective our approach is in terms of automatically detecting and classifying vulnerabilities in IoT networks.

In Table II, we illustrate in detail the comparison with other techniques established for the classification of vulnerability. Our method is compared with some state-of-the-art methods. Kumar et al. [38] proposed a new deep CNN model, Coyote Optimization-based Deep CNN, which obtained, 95% accuracy in vulnerability classification. Li et al. [42] designed an integrated model that stacked multiple CNNs and achieved an accuracy of 86.95%. In 2020, Al-Haija et al. [5] introduced the CNN-ResNet-18 model that performed well with 98.22% accuracy. Varghese et al. [80, 82] used a CNN model trained with MIT-BIH ECG and obtained high accuracy (99.09%). Njima et al. [55] proposed CNNEOS with Word Context (CNNLocWC) and, achieved 94.13%.

Our results showed that the,CNN-ResNet-50 Model performed well in vulnerability classification (96% accuracy). Interestingly, when combined with Tabu Search Optimization, the performance of CNN-Tabu Search was significantly higher than, its constituent, with an accuracy of 99.03%. These results demonstrate that our proposed approach outperforms several state-of-the-art methods and reveal the efficiency of the combination of a CNN architecture with metaheuristic optimization to improve the classification of vulnerability on an IoT network.

There are differing opinions, on both the advantages and disadvantages of vulnerability discovery. The comparisons provide an insight into the, pros and cons of the proposed procedure. Our method, based on, Convolutional Neural Networks (CNN) and Tabu Search Optimization, proved high performance in classifying the vulnerabilities in IoT networks. Allowing Tabu into the search makes it possible for the model to navigate it more effectively and hence leads to higher precision in detecting, vulnerabilities.

However, in spite of the encouraging findings, there are, limitations to our study. For example, despite satisfactory precision of the model, there is still, a danger of overfitting, especially in complex and evolving IoT networks. Furthermore, Tabu Search Optimization could be, more computationally expensive than standard machine learning techniques. So applying the approach in practice demands making a trade-off between the amount of resources and computational complexity.

From this total number, CNN-Tabu search can reach an agreement to classify IoT network vulnerabilities at 96.00% on average. The high accuracy of code similarity analysis and its performance make it possible to promote IoT network security as well as more accurate vulnerability assessment. Nevertheless, its effectiveness over larger and domain-diverse datasets needs to be validated, as well as the possibility of applying it in real scenarios.

TABLE. II. COMPARATIVE ANALYSIS OF ACCURACY IN VULNERABILITY CLASSIFICATION

Study	Model	Accuracy
Kumar et al., 2022	Coy-GWO-based Deep CNN	95.00%
Li et al., 2020	Multi-CNN Fusion Model	86.95%
Al-Haija et al., 2020	CNN-ResNet-18	98.22%
Varghese et al., 2022	CNN MIT-BIH ECG	99.09%
Njima et al., 2019	CNNLocWC	94.13%
Our Study	CNN-Restnet-50	97.00%
Our Study	CNN-Restnet-50 + Tabu Search	99.03%

5) Visualizing results and model interpretability: The confusion matrix visualizations are utilized to exhibit a more specific analysis of the vulnerability classification performance of the proposed EVCIN framework. These matrices provide an explicit depiction of the classification results, which helps to assess how well the model can differentiate individual categories in a vulnerability-based distribution.

The rows in a 'confusion matrix' are the actual classes, and the columns are the predicted classes. Each cell shows the number of individuals for a given class and its prediction. This representation gives a well-founded understanding of how good the model in question is when it comes to distinguishing between vulnerabilities from all categories.

According to the analysis of the Confusion Matrix, EVCIN has proved to have a very high performance in vulnerability identification. The performance is impressive for the DDoS, ICMP, and DDoS: UDP, with both precision, recall, and F1-score being 1.00 (accuracy = 0.991). This ability to correctly recognize the vulnerabilities is also evidenced by its high performance measures on the MITM, DDoS_ TCP Vulnerability scanner, and Normal classes, which include accuracy, precision, recall, and F1-scores. However, the Confusion Matrix also presents certain difficulties in classification. Namely, the fingerprinting, DDoS_HTTP, and XSS classes show reduced precision, recall, and F1-scores with adequate room for improvement in the detection and discrimination of these vulnerabilities. In addition, the average precision, recall, and F1-score of the model are calculated at global, micro, and weighted levels, which assess its goodness. These metrics take account of the class imbalance in the dataset as explained in Fig. 11 and thus provide a complete perspective on the classification performance of the model.

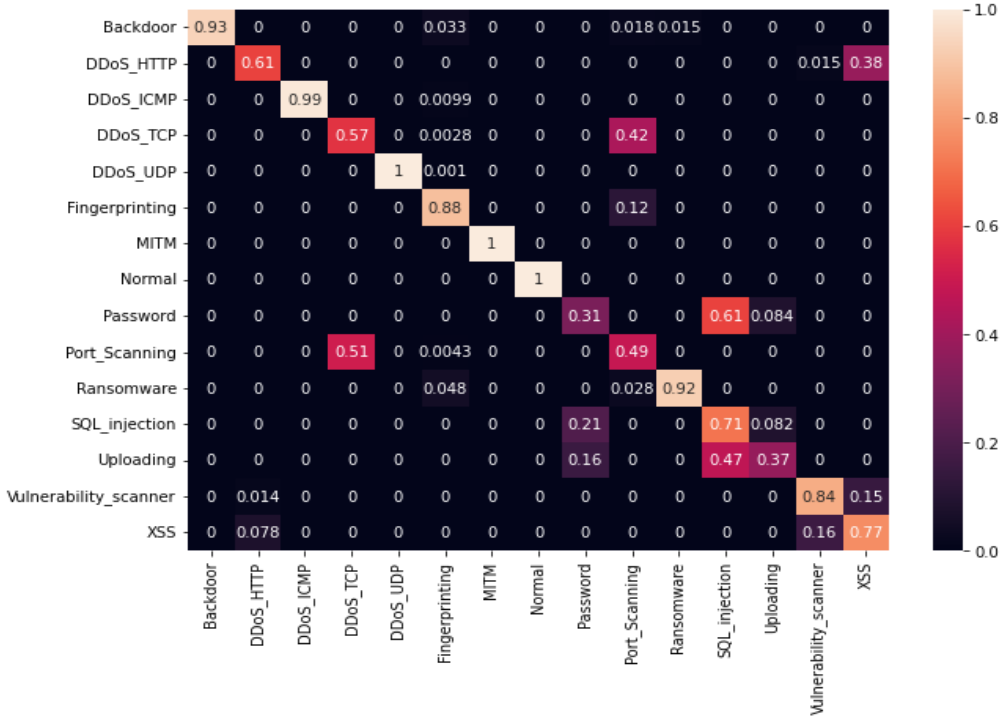


Fig. 11. Illustration of the confusion matrix.

To sum up, Confusion Matrix representations provide interesting points of view regarding the EVCIN framework for vulnerability classification. The good performance of the model is indicated by high accuracy and well-balanced precision, recall, and F1 scores between various classes in the detection and classification of vulnerabilities. Besides, these visualizations are useful to identify the particular categories that may need more exploration, guiding us on how to improve the model performance overall to secure IoT networks.

In order to have a reliable and trustworthy EVCIN framework, it is necessary to comprehend the pivotal role of various features in vulnerability detection. We used a variety of methods to explain feature importance, such as CNN-based feature importance, scores, SHAP, LIME and Grad-CAM. The results have been plotted in Fig. 12, which shows that the network traffic pattern, packet size, and payload contents are significant for vulnerability classification. By locating both false negatives and false positives, the accuracy of classification is also enhanced. By revealing how the model works, people (including network engineers and security analysts) have a clearer view to take action on detected weaknesses. The inclusion of feature importance analysis makes the EVCIN framework more effective and robust in its ability to intelligently identify vulnerabilities in IoT networks.

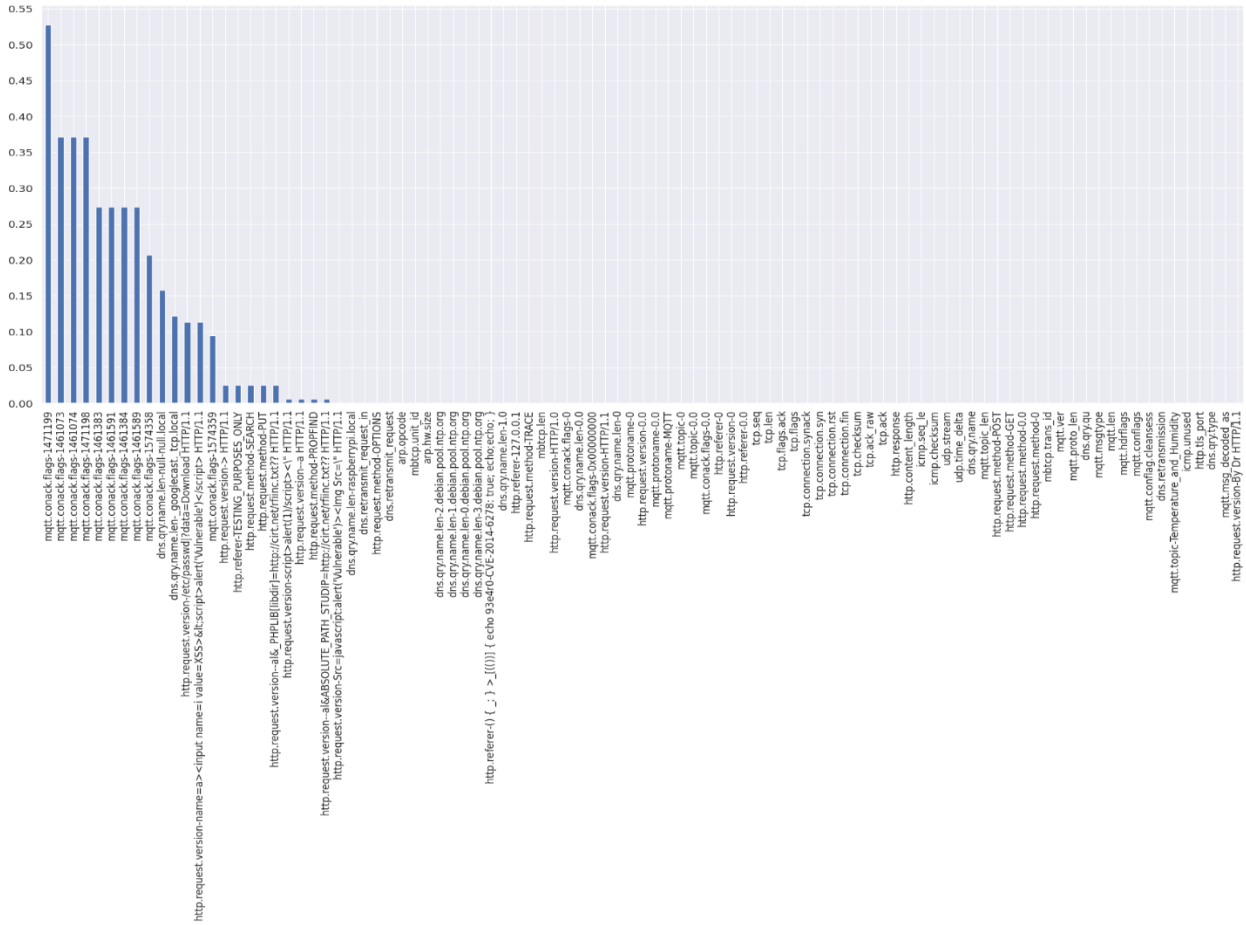


Fig. 12. Illustration of feature selection.

Moreover, in qualitatively evaluating our EVCIN framework for vulnerability identification, we conducted ROC curves and AUC analyses comparison, which is depicted in Fig. 13. These plots provide a vivid portrayal of the model's True Positive Rate (Recall) vs. () * False Positive Rate with different decision threshold values. ROC curves offer a graphical view of the trade-off between true positive detection and false positive minimization. Algorithms with a high sensitivity paired with a low false positive rate will show curves that are slanted towards the upper-left corner of the graph, indicating good performance.

The general, performance of the model can also be evaluated with AUC (Area Under the Curve). An AUC value close to 1 implies that the model can well discriminate different classes, of vulnerabilities. Through the ROC curves together with their AUC metrics, it is evident that our EVCIN framework provides a good vulnerability classification performance, showing qualitative robustness and precision. Analyzing these visual representations becomes a key tool to determine appropriate decision thresholds as part of practical deployment in the Internet, of Things (IoT) domain.

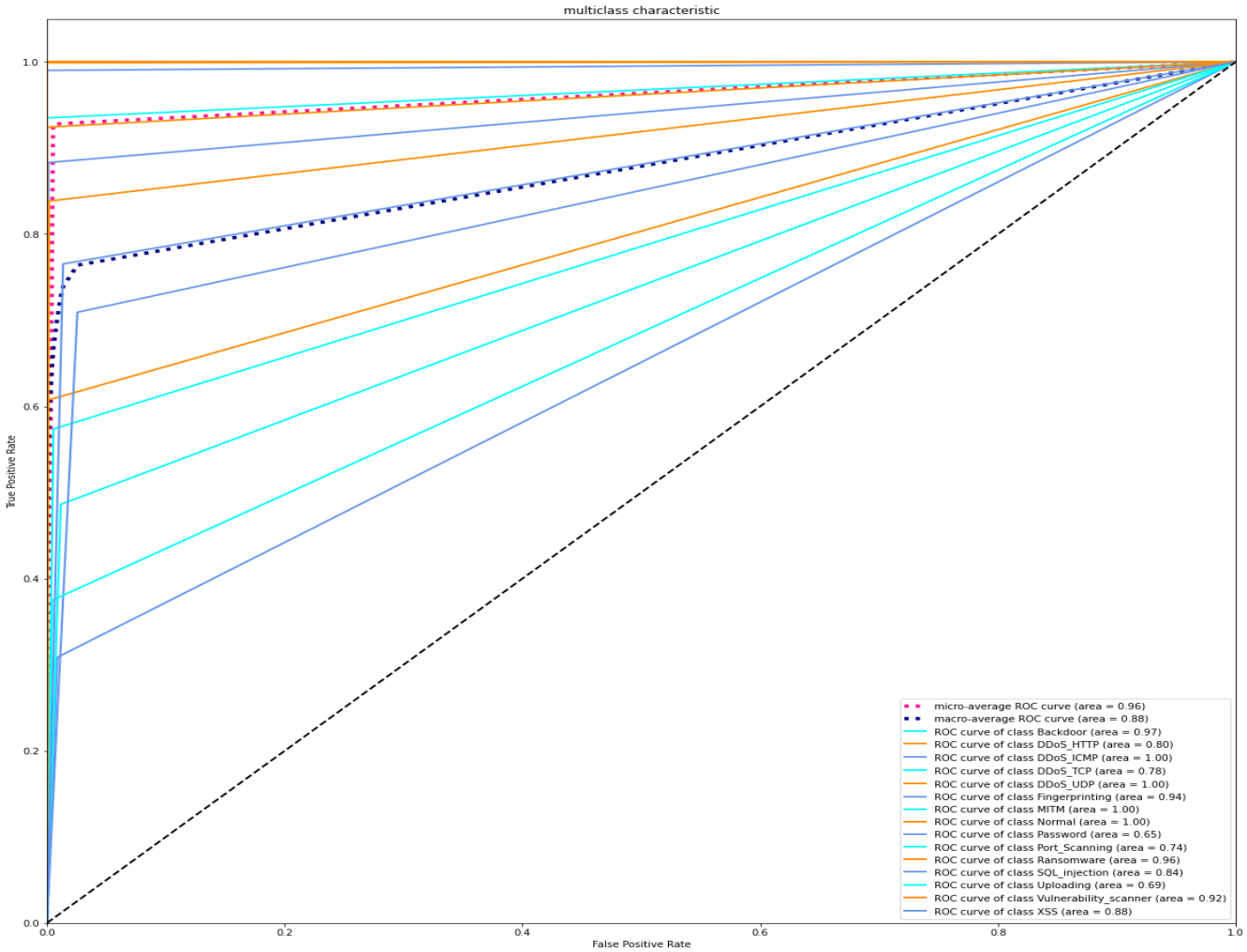


Fig. 13. Graphical representations of ROC curves and AUC analysis.

V. Discussion and Implications of Findings

This section provides an analysis and discussion of the results produced by our CNN-Tabu Search Optimization framework released in relation to vulnerability type classification in IoT networks. Finally, we discuss the relevance of these results in terms of vulnerability discovery and classification for IoT contexts. Furthermore, comparative performance analysis is also performed, demonstrating the merits and, importance of using CNN architecture in Tabu Search Optimization.

In this research, we evaluated the effectiveness of CNN-Tabu Search-based vulnerability classification in comparison with some widely used methods. Our analysis comprised cutting-edge implementation of both Convolutional Neural Networks(CNN) and Tabu Search, which are popular techniques in the research community. Results from performance metrics such as accuracy, precision, and F1 Score indicated that the combined CNN-Tabu Search approach performed better than, other models of CNN and Tabu Search. This favorable relation also may be due to the combined strengths of, both methods: CNNs™ feature extraction capability and Tabu Search's efficiency in optimizing the model parameters. Through the integrated model, which fuses these methods in a unified structure, relatively better performance was achieved for vulnerability identification and classification, indicating the benefits of using different sensing techniques when enhancing IoT network security.

Our CNN-Tabu Search approach performs well in addressing the challenges of vulnerability classification in IoT networks. With the constantly changing nature of IoT network traffic, attackers are forever adjusting their tactics to take advantage of things they should, not do. Based on the net using Convolutional Neural Network and Tabu Search optimization, this study extracts vivid feature representations and learns deep patterns that, reflect different vulnerability types. By combining with Tabu Search algorithms, the optimization process becomes more effective in the search space and converges faster and more reliable, toward optimum solution. As a result, the technique actually increases the accuracy of vulnerability classification, especially for the rare or most difficult vulnerabilities.

VI. Conclusion

This study introduces a novel approach for accurately classifying vulnerabilities in Internet of Things (IoT) networks

by combining Convolutional Neural Networks (CNNs) with Tabu Search Optimization. The goal was to tackle the ever-evolving challenges posed by cyberattacks in IoT environments. To strengthen the security and resilience of IoT infrastructures, our research focused on improving both the accuracy and efficiency of vulnerability detection and classification. Through extensive experiments and evaluations, we demonstrated the effectiveness of the CNN-Tabu Search approach in identifying and categorizing vulnerabilities within IoT network traffic. The use of Convolutional Neural Networks enabled efficient feature extraction and representation, allowing the model to recognize complex patterns indicative of various vulnerabilities. At the same time, Tabu Search Optimization introduced strategic mechanisms for diversification and intensification, which enhanced model accuracy and accelerated convergence toward optimal solutions. A comparative analysis with other state-of-the-art methods, including standalone CNN and Tabu Search models, confirmed the superior performance of our integrated approach. Across key evaluation metrics such as Accuracy, Precision, Recall, and F1 Score, the CNN-Tabu Search framework consistently outperformed alternatives, demonstrating its effectiveness in classifying IoT network vulnerabilities and highlighting the value of combining deep learning with optimization techniques in a unified framework.

Additionally, the interpretability of our model provided security analysts with valuable insights into its decision-making process, enhancing their understanding of the model and increasing confidence in its predictions. In mission-critical IoT applications, transparency in security decisions is essential, making interpretability a key feature. The implications of our findings are highly relevant to cybersecurity in IoT networks. The CNN-Tabu Search approach not only helps security professionals efficiently mitigate risks but also advances the field of vulnerability classification. By accurately identifying and categorizing vulnerabilities in real-time IoT traffic, the proposed methodology strengthens the security of IoT infrastructures and reduces exposure to complex, emerging cyber threats. Our results highlight the potential of integrating advanced deep learning techniques with optimization algorithms to tackle the unique security challenges posed by the rapidly expanding IoT ecosystem. Overall, the CNN-Tabu Search framework demonstrates considerable promise for enhancing cybersecurity research and practice, emphasizing the importance of innovative, interdisciplinary approaches to protecting IoT infrastructures in an increasingly dynamic and evolving landscape.

ACKNOWLEDGMENT

We would like to convey our sincere gratitude to everyone who helped and contributed to completing this research. They have been incredibly helpful and supportive in making this study possible.

REFERENCES

[1] O. I. Abiodun, E. O. Abiodun, M. Alawida, R. S. Alkhawaldeh, and H. Arshad, "A review on the Internet of Things security: Challenges and solutions," Wireless Personal Communications, vol. 119, no. 3, pp. 2603–2637, 2021.

[2] S. M. Abir, S. N. Islam, A. Anwar, A. N. Mahmood, and A. M. Oo, "Building resilience against covid-19 pandemic using artificial intelligence, Machine Learning, and IOT: A survey of recent progress," IoT, vol. 1, no. 2, pp. 506–528, 2020.

[3] G. Abraham, R. Raksha, and M. Nithya, "Smart agriculture based on IoT and machine learning," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 414-419, IEEE, April 2021.

[4] S. F. Ahmed, Md. S. Alam, M. Hoque, A. Lameesa, S. Afrin, T. Farah, M. Kabir, G. Shafiullah, and S. M. Muyeen, "The Industrial Internet of Things enabled technologies, challenges, and Future Directions," Computers and Electrical Engineering, vol. 110, article 108847, 2023.

[5] Q. A. Al-Haija, M. A. Smadi, and S. Zein-Sabatto, "Multi-class weather classification using ResNet-18 CNN for autonomous IoT and CPS applications," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1586-1591, IEEE, December 2020.

[6] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130-165150, 2020.

[7] Z. S. I. Ameen, A. S. Mubarak, C. Altrjman, S. Alturjman, and R. A. Abdulkadir, "Explainable Residual Network for Tuberculosis Classification in the IoT Era," in 2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era (FoNeS-AIoT), pp. 9-12, IEEE, December 2021.

[8] H. K. Apat, R. Nayak, and B. Sahoo, "A comprehensive review of Internet of Things application placement in the fog computing environment," Internet of Things, vol. 23, article 100866, 2023.

[9] S. Arisdakessian, O. A. Wahab, A. Mourad, H. Otrok, and M. Guizani, "A survey on IOT intrusion detection: Federated Learning, game theory, social psychology, and explainable AI as future directions," IEEE Internet of Things Journal, vol. 10, no. 5, pp. 4059–4092, 2022.

[10] M. Asam, S. H. Khan, A. Akbar, S. Bibi, T. Jamal, A. Khan, and M. R. Bhutta, "IoT malware detection architecture using a novel channel boosted and squeezed CNN," Scientific Reports, vol. 12, no. 1, article 15498, 2022.

[11] O. I. Abiodun, E. O. Abiodun, M. Alawida, R. S. Alkhawaldeh, and H. Arshad, "A review on the Internet of Things security: Challenges and solutions," Wireless Personal Communications, vol. 119, no. 3, pp. 2603–2637, 2021.

[12] S. M. Abir, S. N. Islam, A. Anwar, A. N. Mahmood, and A. M. Oo, "Building resilience against covid-19 pandemic using artificial intelligence, Machine Learning, and IOT: A survey of recent progress," IoT, vol. 1, no. 2, pp. 506–528, 2020.

[13] G. Abraham, R. Raksha, and M. Nithya, "Smart agriculture based on IoT and machine learning," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 414-419, IEEE, April 2021.

[14] S. F. Ahmed, Md. S. Alam, M. Hoque, A. Lameesa, S. Afrin, T. Farah, M. Kabir, G. Shafiullah, and S. M. Muyeen, "The Industrial Internet of Things enabled technologies, challenges, and Future Directions," Computers and Electrical Engineering, vol. 110, article 108847, 2023.

[15] H. Wang, Y. Zhao, and X. Liu, "A Secure and Efficient Data Transmission Scheme Based on Blockchain in Internet of Medical Things," Wireless Communications and Mobile Computing, vol. 2021, Article ID 5047355, 2021. doi: 10.1155/2021/5047355.

[16] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130-165150, 2020.

[17] Z. S. I. Ameen, A. S. Mubarak, C. Altrjman, S. Alturjman, and R. A. Abdulkadir, "Explainable Residual Network for Tuberculosis Classification in the IoT Era," in 2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era (FoNeS-AIoT), pp. 9-12, IEEE, December 2021.

[18] H. K. Apat, R. Nayak, and B. Sahoo, "A comprehensive review of Internet of Things application placement in the fog computing environment," Internet of Things, vol. 23, article 100866, 2023.

[19] C. Roy, S. S. Yadav, V. Pal, M. Singh, S. K. Patra, and G. R. Sinha,

"[Retracted] An Ensemble Deep Learning Model for Automatic Modulation Classification in 5G and Beyond IoT Networks," Computational Intelligence and Neuroscience, vol. 2021, Article ID 5047355, 8 pages, 2021. doi: 10.1155/2021/5047355.

[20] M. Asam, S. H. Khan, A. Akbar, S. Bibi, T. Jamal, A. Khan, and M. R. Bhutta, "IoT malware detection architecture using a novel channel boosted and squeezed CNN," Scientific Reports, vol. 12, no. 1, article 15498, 2022.

[21] M. Cheminod, L. Durante, and A. Valenzano, "Review of security issues in Industrial Networks," IEEE Transactions on Industrial Informatics, vol. 9, no. 1, pp. 277-293, 2013.

[22] A. Dahou et al., "Intrusion detection system for IOT based on Deep Learning and modified reptile search algorithm," Computational Intelligence and Neuroscience, vol. 2022, pp. 1-15, 2022.

[23] G. K. Dasebenezer and B. Joselin, "TSO clustered protocol to extend lifetime of IOT based Mobile Wireless Sensor Networks," The International Arab Journal of Information Technology, vol. 20, no. 4, 2023.

[24] N. Denis et al., "Privacy-preserving content-based publish/subscribe with encrypted matching and data splitting," in SECRYPT 2020: 17th International Conference on Security and Cryptography, 2020, pp. 405-414.

[25] S. L. Ullo and G. R. Sinha, "Advances in IoT and Smart Sensors for Remote Sensing and Agriculture Applications," Remote Sensing, vol. 13, no. 13, p. 2585, 2021. doi: 10.3390/rs13132585.

[26] A. Ed-daoudy and K. Maalmi, "A new internet of things architecture for real-time prediction of various diseases using machine learning on Big Data Environment," Journal of Big Data, vol. 6, no. 1, 2019.

[27] A. B. Farid et al., "Software defect prediction using hybrid model (CBIL) of Convolutional Neural Network (CNN) and bidirectional long short-term memory (Bi-LSTM)," PeerJ Computer Science, vol. 7, 2021.

[28] S. Galiveeti et al., "Cybersecurity analysis: Investigating the data integrity and privacy in AWS and Azure Cloud Platforms," Studies in Big Data, pp. 329-360, 2021.

[29] Y. Guo et al., "Deep Learning for Visual Understanding: A Review," Neurocomputing, vol. 187, pp. 27-48, 2016.

[30] A. U. Haq et al., "DACBT: Deep learning approach for classification of brain tumors using MRI data in IoT healthcare environment," Scientific Reports, vol. 12, no. 1, p. 15331, 2022.

[31] A. U. Haq et al., "MCNN: a multi-level CNN model for the classification of brain tumors in IoT-healthcare system," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 5, pp. 4695-4706, 2023.

[32] D. Kumar and M. Kumar, "Hybrid Cryptographic Approach for Data Security Using Elliptic Curve Cryptography for IoT," International Journal of Computer Network and Information Security, vol. 16, no. 2, pp. 42-54, 2024. doi: 10.5815/ijcnis.2024.02.04.

[33] A. N. Jahromi et al., "An improved two-hidden-layer extreme learning machine for malware hunting," Computers & Security, vol. 89, p. 101655, 2020.

[34] V. A. Jane, "Survey on iot data preprocessing," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 9, pp. 238-244, 2021.

[35] E. C. Joseph, "Development of Smart IoT-Based CNN Technique for Harmful Maize Insects Recognition in Precision Agriculture," International Journal of Scientific Research in Computer Science and Engineering, vol. 9, pp. 48-60, 2021.

[36] S. Kapoor et al., "A comparative study on Deep Learning and machine learning models for human action recognition in aerial videos," The International Arab Journal of Information Technology, vol. 20, no. 4, 2023.

[37] A. R. Khan et al., "Deep learning for intrusion detection and security of internet of things (IOT): Current analysis, challenges, and possible solutions," Security and Communication Networks, vol. 2022, pp. 1-13, 2022.

[38] A. Kumar et al., "IoT-based ECG monitoring for arrhythmia classification using Coyote Grey Wolf optimization-based deep learning CNN classifier," Biomedical Signal Processing and Control, vol. 76, p. 103638, 2022.

[39] S. Kumar, P. Tiwari, and M. Zymbler, "Internet of things is a revolutionary approach for future technology enhancement: A Review," Journal of Big Data, vol. 6, no. 1, 2019.

[40] A. Lavric and V. Popa, "Internet of things and LoraTM low-power wide-area networks: A survey," in 2017 International Symposium on Signals, Circuits and Systems (ISSCS), 2017.

[41] I. Lee, "Internet of things (IOT) cybersecurity: Literature review and IOT cyber risk management," Future Internet, vol. 12, no. 9, p. 157, 2020.

[42] Y. Li et al., "Robust detection for network intrusion of industrial IoT based on multi-CNN fusion," Measurement, vol. 154, p. 107450, 2020.

[43] Z. Liu et al., "Using embedded feature selection and CNN for classification on CCD-INID-V1—a new IoT dataset," Sensors, vol. 21, no. 14, p. 4834, 2021.

[44] D. B. Darshan and P. C. R., "Dual-discriminator Conditional Generative Adversarial Network Optimized with Hybrid Momentum Search Algorithm and Giza Pyramids Construction Algorithm for Cluster-Based Routing in WSN Assisted IoT," International Journal of Computer Network and Information Security, vol. 15, no. 5, pp. 96-112, 2023. doi: 10.5815/ijcnis.2023.05.09.

[45] J. Luo et al., "A Secure Transmission Scheme of Sensitive Power Information in Ubiquitous Power IoT," in Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence, 2019.

[46] Y. Ma et al., "Data preprocessing of agricultural IoT based on time series analysis," in Intelligent Computing Theories and Application: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part I, 2018.

[47] J. Manhas and S. Kotwal, "Implementation of intrusion detection system for internet of things using Machine Learning Techniques," Multimedia Security, pp. 217-237, 2021.

[48] A. Marchisio et al., "SEVUC: A study on the security vulnerabilities of capsule networks against adversarial attacks," Microprocessors and Microsystems, vol. 96, p. 104738, 2023.

[49] S. Mishra and A. K. Tyagi, "The role of machine learning techniques in internet of things-based cloud applications," Internet of Things, pp. 105-135, 2022.

[50] V. Mohammadi et al., "Trust-based recommendation systems in internet of things: A systematic literature review," Human-Centric Computing and Information Sciences, vol. 9, no. 1, 2019.

[51] H. Naeem, B. M. Alshammari, and F. Ullah, "Explainable artificial intelligence-based IOT device malware detection mechanism using image visualization and fine-tuned CNN-based Transfer Learning Model," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[52] N. K. Narang, "Mentor's musings on the role of Disruptive Technologies and innovation in making healthcare systems more sustainable," IEEE Internet of Things Magazine, vol. 4, no. 3, pp. 80-89, 2021.

[53] R. A. S. Naseri, A. Kurnaz, and H. M. Farhan, "Optimized face detector-based intelligent face mask detection model in IoT using deep learning approach," Applied Soft Computing, vol. 134, p. 109933, 2023.

[54] N. Neshenko et al., "Demystifying IOT security: An exhaustive survey on IOT vulnerabilities and a first empirical look on internet-scale IOT exploitations," IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2702-2733, 2019. doi: 10.1109/comst.2019.2910750.

[55] W. Njima et al., "Deep CNN for indoor localization in IoT-sensor systems," Sensors, vol. 19, no. 14, p. 3127, 2019.

[56] B. Nour and S. Cherkaoui, "Unsupervised Data Splitting Scheme for Federated Edge Learning in IoT Networks," in ICC 2022-IEEE International Conference on Communications, 2022, pp. 1-6.

[57] P. R. Maidamwar, P. P. Lokulwar, and K. Kumar, "Ensemble Learning Approach for Classification of Network Intrusion Detection in IoT Environment," International Journal of Computer Network and Information Security, vol. 15, no. 3, pp. 30-46, 2023. doi: 10.5815/ijcnis.2023.03.03.

[58] S. Ojagh et al., "Enhanced air quality prediction by edge-based spatiotemporal data preprocessing," Computers & Electrical Engineering, vol. 96, p. 107572, 2021.

[59] I. Omara et al., "A hybrid model combining learning distance metric and DAG support Vector Machine for multimodal biometric recognition," IEEE Access, vol. 9, pp. 4784-4796, 2021. doi: 10.1109/access.2020.3035110.

[60] M. Osman et al., "ML-LGBM: A machine learning model based on light gradient boosting machine for the detection of version number attacks in RPL-based networks," IEEE Access, vol. 9, pp. 83654-83665, 2021. doi: 10.1109/access.2021.3087175.

[61] M. Patel and M. Bhise, "Raw data processing framework for IoT," in 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 2019, pp. 695-699. doi: 10.1109/comsnet.2019.8711477.

[62] S. K. Peddoju and H. Upadhyay, "Evaluation of IoT data visualization tools and techniques," in Data visualization: Trends and challenges toward multidisciplinary perception, 2020, pp. 115-139.

[63] V. Ponnusamy and S. Natarajan, "Precision agriculture using advanced technology of IoT, unmanned aerial vehicle, augmented reality, and machine learning," in Smart Sensors for Industrial Internet of Things: Challenges, Solutions and Applications, 2021, pp. 207-229.

[64] M. S. Pour et al., "Data-driven curation, learning and analysis for inferring evolving iot botnets in the wild," in Proceedings of the 14th International Conference on Availability, Reliability and Security, 2019, pp. 1-10.

[65] D. Puthal et al., "Decision tree based user-centric security solution for critical IOT infrastructure," Computers and Electrical Engineering, vol. 99, p. 107754, 2022. doi: 10.1016/j.compeleceng.2022.107754.

[66] W. Rafique et al., "Complementing IOT services through software defined networking and edge computing: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1761-1804, 2020. doi: 10.1109/comst.2020.2997475.

[67] S. Rizvi et al., "Threat model for securing internet of things (IOT) network at device-level," Internet of Things, vol. 11, p. 100240, 2020. doi: 10.1016/j.iot.2020.100240.

[68] P. D. Rosero-Montalvo et al., "A new data-preprocessing-related taxonomy of sensors for iot applications," Information, vol. 13, no. 5, p. 241, 2022.

[69] T. Saba et al., "Intrusion detection system through advance machine learning for the internet of things networks," IT Professional, vol. 23, no. 2, pp. 58-64, 2021. doi: 10.1109/mitp.2020.2992710.

[70] T. J. Saleem and M. A. Chishti, "Deep learning for the internet of things: Potential benefits and use-cases," Digital Communications and Networks, vol. 7, no. 4, pp. 526-542, 2021. doi: 10.1016/j.dcan.2020.12.002.

[71] K. Sha et al., "A detailed review of implementation of Deep Learning Approaches for Industrial Internet of things with the different opportunities and challenges," in 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), 2022, pp. 1-18. doi: 10.1109/ic3i56241.2022.10072499.

[72] K. Sha et al., "A survey of Edge Computing-based designs for IOT Security," Digital Communications and Networks, vol. 6, no. 2, pp. 195-202, 2020. doi: 10.1016/j.dcan.2019.08.006.

[73] E. A. Shammar et al., "A survey of IOT and Blockchain Integration: Security perspective," IEEE Access, vol. 9, pp. 156114-156150, 2021. doi: 10.1109/access.2021.3129697.

[74] A. Shamsoshoara et al., "A survey on physical unclonable function (puf)-based security solutions for internet of things," Computer Networks, vol. 183, p. 107593, 2020. doi: 10.1016/j.comnet.2020.107593.

[75] C. Shao et al., "IoT data visualization for business intelligence in corporate finance," Information Processing & Management, vol. 59, no. 1, p. 102736, 2022.

[76] O. Sharma et al., "Windows and IoT malware visualization and classification with deep CNN and Xception CNN using Markov images," Journal of Intelligent Information Systems, vol. 60, no. 2, pp. 349-375, 2023.

[77] S. K. Singh et al., "Machine learning based distributed big data analysis framework for next generation web in IoT," Computer Science and Information Systems, vol. 18, no. 2, pp. 597-618, 2021.

[78] M. Soori et al., "Internet of things for smart factories in Industry 4.0, a review," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 192-204, 2023. doi: 10.1016/j.iotcps.2023.04.006.

[79] S. Sun et al., "Integrating Network Function Virtualization with SDR and SDN for 4G/5G networks," IEEE Network, vol. 29, no. 3, pp. 54-59, 2015. doi: 10.1109/mnet.2015.7113226.

[80] C. S. Suresh et al., "Cognitive IoT-based smart fitness diagnosis and recommendation system using a three-dimensional CNN with hierarchical particle swarm optimization," in Smart Sensors for Industrial Internet of Things: Challenges, Solutions and Applications, 2021, pp. 147-160.

[81] G. Tsaramirsis et al., "A modern approach towards an industry 4.0 model: From Driving Technologies to management," Journal of Sensors, vol. 2022, p. 5023011, 2022. doi: 10.1155/2022/5023011.

[82] A. Varghese et al., "Conception and realization of an IoT-enabled deep CNN decision support system for automated arrhythmia classification," Journal of Intelligent Systems, vol. 31, no. 1, pp. 407-419, 2022.

[83] Y. Wang et al., "Image source identification using convolutional neural networks in IoT environment," Wireless Communications and Mobile Computing, vol. 2021, p. 5790730, 2021.

[84] Mohammad Othman Nassar, Feras Fares AL-Mashagba, "FED-SCADA: A Trustworthy and Energy-efficient Federated IDS for Smart Grid Edge Gateways Using SNNs and Differential Evolution", International Journal of Computer Network and Information Security (IJCNIS), vol. 17, no. 6, pp. 1–15, December. 2025, doi: https://doi.org/10.5815/ijcnis.2025.06.01

[85] M. Zolanvari et al., "Machine learning-based network vulnerability analysis of Industrial Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6822-6834, 2019. doi: 10.1109/jiot.2019.2912022.

[86] M. Zorzi et al., "From today's Intranet of things to a future internet of things: A wireless- and mobility-related view," IEEE Wireless Communications, vol. 17, no. 6, pp. 44-51, 2010. doi: 10.1109/mwc.2010.5675777.

[87] Nassar, M. O., & Al-Mashagba, F. F. (2025). Optimal ensemble learning with meta-heuristics for multiclass classification of syscall-binder interactions in mobile applications. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 16(1), 26–48. https://doi.org/10.58346/JOWUA.2025.I1.002

[88] Al Ghamri, M., Ibrahim, D., Sihwail, R., & Shehab, M. (2024). Whale Optimization Algorithm for Feature Selection Enhances Classification in Malware Datasets. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE42024233

[89] Ibrahim, D., Sihwail, R., Arrifin, K. A. Z., Abuthawabeh, A., & Mizher, M. (2023). A Novel Color Visual Cryptography Approach Based on Harris Hawks Optimization Algorithm. *Symmetry*, 15(7), 1305. https://doi.org/10.3390/sym15071305

[90] Shannaq, F., Shehab, M., Alshorman, A., Hammad, M., Hammo, B., & Al-Omari, W. (2025). Exploring Metaheuristic Optimization Algorithms in the Context of Textual Cyberharassment: A Systematic Review. *Expert Systems*, 42, e13826. https://doi.org/10.1111/exsy.13826

[91] Daoud, M. S., Shehab, M., Abualigah, L., et al. (2023). Hybrid Modified Chimp Optimization Algorithm and Reinforcement Learning for Global Numeric Optimization. *Journal of Bionic Engineering*, 20, 2896–2915. https://doi.org/10.1007/s42235-023-00394-2

AUTHORS' PROFILES



Dr. Feras almashakbah is an Associate Professor of Artificial Intelligence in Jerash University; he received his Ph.D. in Computer Information System/ Artificial Intelligence in 2009. Dr. Almashakbah has 15 years' teaching experience. He has published 30 scientific research publications in international peer reviewed journals and conferences. Dr. Almashakbah is an editorial board member for 3 international peer reviewed journals.



Dr. Mohammed Nassar is an Associate Professor of Computer Information System in Amman Arab University; he received his Ph.D. in Computer Information System/ Information Retrieval in 2009. Dr. Nassar has 15 years' teaching experience. He has been working at Amman Arab University since 2010. His is the Manager of eLearning center and the director of marketing department in Amman Arab University. He occupied different leading positions at the university: head for Computer Information System department, and computer center manager in Amman Arab University. He has published 40 scientific research publications in international peer reviewed journals and conferences. Dr. Nassar is an editorial board member for 4 international peer reviewed journals.



Dr. Essam S. Hanandeh Associate Professor, Essam Said Hanandeh received the degree of a Ph.D. in CIS in 2008, he joined Zarqa University in Jordan in 2008 up to 2024 then joind to Jerash university in 2025. Associate Professor Essam Said Hanandeh has been worked for 15 years as programmer & System Analyst. He published over 30 research paper in international journals and conferences. He can be contacted at E-mail: e.hanandeh@jpu.edu.jo

Scoping Review on Global Digital Policies and Transformation Strategies for Ageing Societies: Implications for Malaysia

Ahmad Wiraputra Selamat, Shamsul Arrieya Ariffin
Department of Software Engineering and Smart Technology,
Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Perak, Malaysia

***Abstract*—The rapid growth of ageing populations presents major social, healthcare, and economic challenges worldwide. Governments are increasingly adopting digital transformation strategies to support ageing societies through artificial intelligence (AI), the Internet of Things (IoT), telehealth, smart environments, and interoperable data systems. However, existing studies often focus on individual technologies, single-country initiatives, or sector-specific programmes, resulting in a fragmented understanding of how digital policies address ageing populations at the global level. This study conducts a scoping review to map digital policies, frameworks, and transformation strategies related to ageing societies and to identify transferable implications for Malaysia. The review follows PRISMA-ScR guidance and applies a Population-Concept-Context (PCC) framework. Searches were conducted across Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, and institutional repositories for publications and policy documents from 2015 to 2025. After screening 857 records, 63 sources of evidence were included. The findings show that global digital ageing policies cluster around five domains: healthcare digitalisation, social inclusion and active ageing, smart environments and age-friendly cities, governance and data regulation, and economic participation of older adults. Persistent challenges include digital literacy gaps, privacy and ethical concerns, fragmented governance structures, limited interoperability, and unequal digital infrastructure. This review provides a consolidated global overview and highlights policy mechanisms that can inform the development of integrated digital ageing frameworks, particularly for countries preparing for population ageing, such as Malaysia.**

Keywords*—Ageing nation; digital framework; ageing society; digital transformation; scoping review

I. Introduction

The global demographic landscape is undergoing a significant transformation characterized by an unprecedented increase in the ageing population. Projections indicate that by 2030, one in six individuals worldwide will be aged 60 years or over, with the number rising from 1 billion in 2020 to 1.4 billion [1]. By 2050, this demographic is expected to double to 2.1 billion, and the population aged 80 years or older is anticipated to triple, reaching 426 million. This shift presents multifaceted challenges, necessitating comprehensive policies and innovative strategies to ensure the well-being and active participation of older adults in society. The World Health Organization (WHO) has declared 2021-2030 as the United Nations Decade of Healthy Ageing [2].

In response to these demographic changes, digital transformation has emerged as a pivotal approach to address the unique needs of ageing populations. Technologies such as AI, health sensors, and virtual reality applications are being leveraged to enhance healthcare delivery, promote social inclusion, and support independent living among older adults [3].

Over the past decade, numerous countries and regions have introduced digital strategies and policy frameworks aimed at addressing ageing-related challenges. However, the current body of literature remains fragmented and siloed. Much of the research focuses on individual national case studies, specific technologies, or sectoral perspectives, rather than offering a systematic mapping of digital policies and transformation strategies across countries [4]. Consequently, there is limited comparative understanding of how digital transformation is being conceptualized globally in relation to ageing populations, and how different policy approaches address issues such as inclusion, governance, ethics, and sustainability.

Furthermore, critical analyses highlight that many digital ageing policies are characterized by technological optimism, often underestimating the social, ethical, and organizational complexities of ageing and care. This raises concerns about digital exclusion, uneven implementation, and the potential widening of inequalities among older populations if policies are not grounded in integrated, human-centered frameworks [2].

Although several reviews have examined specific technologies or sectoral issues in later life, such as artificial intelligence in elderly healthcare, social housing and ageing in place, and digital health adoption, these works do not comprehensively map national and international digital policies, frameworks, and transformation strategies for ageing societies. Therefore, there remains a need for a scoping review that systematically examines the breadth, characteristics, and conceptual orientations of policy responses across countries and regions.

This scoping review aims to systematically map and synthesize existing digital policies, frameworks, and national digital transformation strategies developed worldwide in response to ageing societies. Specifically, the study seeks to: 1) identify digital policies and frameworks implemented globally to address ageing populations; 2) map the Industry 4.0 and digital components emphasized in these initiatives; 3) compare

policy orientations across regions; and 4) derive implications for Malaysia as it prepares for population ageing. This explicit Malaysia-oriented implication is positioned as an applied extension of the global synthesis rather than as a replacement for the review's international scope.

II. Background

A. Ageing Nation

An ageing nation refers to a demographic shift in which the proportion of older individuals within a country's population surpasses that of younger generations. This phenomenon is primarily driven by three factors: declining fertility rates, increased life expectancy, and advancements in public health systems [5]. According to a study carried out by [6], at least 14 per cent of the population is expected to be aged 65 and above by 2044, with this age group expected to exceed 20 per cent of the population by 2056, giving the country the status of a "super-aged nation".

The shift towards ageing societies is strongly influenced by the demographic legacy of the Baby Boomer generation, generally referring to individuals born during the post-World War II baby boom between 1946 and 1964. Due to its large cohort size, this generation has significantly reshaped population structures as it moves into retirement and older adulthood. This transition has accelerated population ageing in countries such as the United States, Japan, and several European nations, where increased life expectancy and declining fertility have further intensified the ageing process. Similar trends are increasingly evident in developing and upper-middle-income countries, including Malaysia, demonstrating that population ageing is now a global phenomenon with significant implications for healthcare systems, labour markets, social protection, and digital transformation policy [7].

This demographic shift presents multifaceted challenges and opportunities for nations, particularly in areas such as healthcare, economic sustainability, and social inclusion. Ageing societies often face increased healthcare costs, workforce shortages, and a higher dependency ratio, necessitating comprehensive strategies to address these challenges [6]. Concurrently, this transition also presents opportunities for leveraging the potential of older adults in contributing to the economy and society through active ageing policies and digital inclusivity.

B. Digital Policy and Frameworks Around the World

In response to these pressures, digital transformation has been positioned globally as a strategic enabler for supporting ageing societies. Digital health technologies, artificial intelligence (AI), data-driven decision-making, and smart care systems are increasingly promoted as tools to enhance care efficiency, support ageing in place, and improve coordination across service sectors [2]. Digital transformation offers a strategic pathway to address the multifaceted challenges of an ageing population. By leveraging digital technologies, enhancing digital literacy, and fostering collaborative governance, societies can improve the quality of life for older adults and ensure sustainable development [8].

However, scholars caution that digital transformation in ageing contexts is not purely a technical process but a socio-technical transition that reshapes care practices, professional roles, governance arrangements, and citizen-state relationships [9]. Existing academic literature on digital ageing has largely focused on specific domains, particularly eHealth and telemedicine, as mentioned by [9]. Comparative policy analyses, such as studies examining European eHealth strategies, demonstrate that national digital ageing policies differ substantially in their assumptions, governance structures, and emphasis on inclusion and care responsibility [4].

C. Rationale for a Scoping Review

Given the diversity, complexity, and dispersion of digital policies and transformation strategies related to ageing, a scoping review is an appropriate methodological approach. Scoping reviews are particularly suited to mapping broad bodies of evidence, clarifying key concepts, and identifying knowledge gaps in emerging or heterogeneous fields [10].

A preliminary search for existing reviews was conducted using combinations of terms related to "digital ageing", "digital health policy", "eHealth governance", "ageing society", "older adults", "technology adoption", and "scoping review". The search identified reviews focusing on narrower domains, including AI in elderly healthcare, digital health solutions for older adults, social housing, and ageing-in-place services. However, no prior review was found that systematically mapped global digital policies, frameworks, and transformation strategies for ageing societies across multiple continents. This justifies the use of a scoping review to synthesise policy trends, identify dominant conceptual approaches, and highlight gaps for future empirical and policy-oriented research [11],[12].

III. Methodology

This study adopted a scoping review (ScR) approach to systematically map the existing literature on digital policies, frameworks, and transformation strategies related to ageing nations. The review followed the methodological framework of [10] further refined by [13], and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [14]. The use of a scoping review was deemed appropriate given the objectives of this study: 1) to identify the breadth of available evidence, 2) to map key concepts and policies, and 3) to highlight knowledge gaps for future research [15]. A scoping review protocol was developed a priori to guide the review objectives, eligibility criteria, search strategy, and data charting process. However, the protocol was not registered in an international database. Any deviations from the initial protocol were documented and transparently reported in accordance with PRISMA-ScR recommendations [14].

A. Identifying the Research Question and Objectives

The research questions were formulated using the PCC (Population–Concept–Context) framework recommended for scoping reviews [16]:

- Population: Ageing populations and older adults (aged 60+).

- Concept: Digital transformation strategies, policies, and frameworks leveraging digitalisation, especially Industry 4.0 components (AI, IoT, Big Data, etc.).
- Context: National and international responses to ageing nation challenges.

Guided by the Population–Concept–Context (PCC) framework, this scoping review addresses the following research questions:

- What digital policies and frameworks have been implemented globally to address ageing populations?
- What digital and Industry 4.0 components are emphasised in these initiatives?
- What recurring challenges and gaps are reported in the digital transformation of ageing societies?

B. Information Sources and Search Strategy

A comprehensive search was conducted across Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and PubMed. Complementary searches were also performed through institutional repositories and official websites of international organisations and government agencies, including the World Health Organisation (WHO), United Nations (UN), European Union (EU), OECD, national health ministries, digital government agencies, and ageing-related policy portals. The final search was completed in March 2026. The search strategy combined terms related to ageing populations, digital transformation, digital policy, digital health, Industry 4.0, and policy frameworks. Search limits were applied to English-language sources published between 2015 and 2025 to reflect the Industry 4.0 and digital government era.

- ("digital policy" OR "digital governance" OR "e-government") AND ("ageing population" OR "elderly" OR "senior citizens")
- ("digital transformation strategy" OR "smart ageing" OR "technology adoption") AND ("ageing nation" OR "elderly support")
- ("framework" OR "model" OR "guideline") AND ("digital inclusion" OR "elderly technology adoption")

The search was limited to sources published between 2015 and 2025 to ensure coverage of recent digital transformation, Industry 4.0, and post-pandemic digital health developments.

C. Eligibility Criteria

Studies were included if they:

- Sources published between 2015 and 2025, reflecting the Industry 4.0 and digital government era.
- Peer-reviewed journal articles, reputable news reports, conference proceedings, or official policy reviews.
- Focused on digital transformation, Industry 4.0, or policy frameworks in relation to ageing nations.
- English-language publications.

Excluded studies were those that:

- Focused only on clinical interventions without policy/technology perspectives.
- Were editorials, commentaries, or non-peer-reviewed reports.
- Were duplicate records or outside the scope (e.g., rural housing without digital transformation).
- Non-English publications.

D. Study Selection

All records were exported into Mendeley for reference management and duplicate removal. Title and abstract screening were performed independently by two reviewers using the eligibility criteria. Potentially relevant records were then assessed through full-text screening. Disagreements were resolved through discussion until a consensus was reached. Where a source was a policy document rather than an empirical study, the reviewers assessed whether it explicitly addressed ageing, older adults, digital transformation, digital inclusion, digital health, or related governance mechanisms.

E. Data Charting

A structured data-charting form was developed based on scoping review recommendations and piloted on ten sources before full extraction. Minor refinements were made to improve consistency in recording policy type, digital components, and reported implementation challenges. Data charting was conducted by the first reviewer and checked by the second reviewer for completeness and consistency. Extracted data included:

- Author(s), year, country/region
- Study aim and design
- Population focus
- Policy, framework, or strategy discussed
- Digital/Industry 4.0 components addressed
- Reported challenges

F. Data Analysis and Synthesis

Data were synthesised descriptively and presented in narrative, tabular, and visual forms. First, the included sources were grouped by continent, country, or region, and policy type. Second, digital components were coded into technology categories such as AI, IoT, telehealth, big data, smart homes, digital identity, interoperability, and digital literacy programmes. Third, thematic synthesis was used to identify recurring policy domains. Five themes were finalised through reviewer discussion: healthcare digitalisation and smart health; social inclusion and active ageing; smart environments and age-friendly cities; governance, protection, and data regulation; and work, economics, and productivity. This process enabled both descriptive mapping and comparative interpretation across regions.[14]

IV. RESULTS

A total of 801 records were identified through database searches, and an additional 56 from manual screening of references and relevant policy documents, so the total is 857. After removal of 230 duplicates and 2 for other reasons, 625 unique records were screened by title and abstract. Following the exclusion of 343 unrelated studies, 282 full-text papers were reviewed for eligibility. Following full-text assessment, 219 studies were excluded for reasons such as:

- Reason 1: Focused solely on medical/clinical interventions without a policy or technology framework (n = 134)
- Reason 2: Non-peer-reviewed or grey literature (n = 41)
- Reason 3: Duplicated or incomplete text (n = 42)
- Reason 4: Not available in English (n = 2)

A total of 63 sources of evidence were included in the scoping review as summarized in Table I. Selection process following the PRISMA-ScR guideline is illustrated in Fig. 1. The included documents were predominantly from Asia (38%), followed by Europe (27%), Oceania (7%), North America (4%), Africa (4%), South America (5%), and international or global frameworks (4%), as shown in Fig. 2.

TABLE I. SUMMARY OF SOURCES OF EVIDENCE INCLUDED IN THE SCOPING REVIEW

Continent	Country	Sources
Asia	Japan	Society 5.0 for Super-Aged Society [17] Society 5.0 Japan's National Digital Transformation Vision [18], [19] Japan Revitalization Strategy / Future Investment Strategy (2016–2020) [20] The National Framework for Promotion of Dementia Policies (2019) [21] Community-Based Integrated Care System [22]
	Singapore	Smart Nation 2030 [23] Smart Nation Strategy (2018) [23] Seniors Go Digital Program (2020) [24] Singapore AI Home Healthcare for Ageing Populations [25]
	China	Healthy China initiative 2019 – 2030 [26] eHealth for elderly [26]
	Korea	Digital Inclusion Act 2019 [27] National Responsibility for Dementia System [28] 50+ Site and National Informatization Education Site [29]
	Thailand	Thailand 4.0 [30] National Action Plan for Older Persons Phase 3 (2023-2037) [31] Digital Citizen Course by ETDA [32] Digital Economy and Society Development Plan Digital Thailand [33]
	Qatar	National Health Strategy [34]
	United Arab Emirates	National Policy for Senior Emiratis [35]
	India	Digital India Programme – Digital services for seniors [36]
	Indonesia	Digital Indonesia Vision 2045 [37]
	Malaysia	Malaysia Digital Economy Blueprint [1] National Ageing Blueprint [1]
Europe	EU	Green Paper on Ageing: Fostering Solidarity & Responsibility Between Generations (2021) [38] European Innovation Partnership on Active and Healthy Ageing (EIP on AHA) [39] Active and Assisted Living (AAL) Programme – "Ageing Well in the Digital World" [40] EU Artificial Intelligence Act 2024 [41]
	Finland	eHealth [29] Finland's Kanta Services [42]
	Sweden	Effektiv vård (Efficient Care) 2016 [43] Vision eHealth 2025 [44]
	Italy	National Recovery and Resilience Plan (PNRR) [45]
	France	"Ma Santé 2022" – My Health 2022 (Digital Health Component) [46]
	Germany	Digital Healthcare Act (DVG) (2019) [47] Digital Act (DigiG) (2024) [48]
	Russia	Strategy for the Development of an Information Society (2017–2030)[49]
	Denmark	National eHealth Strategy [50]
	United Kingdom	Ageing Society Grand Challenge & Healthy Ageing ISCF [51] NHS Long Term Plan (Digital Transformation for Ageing Care) [52]
	Spain	Espana Digital 2026 [53]
North America	USA	AI-driven elderly care [54] Federal Health IT Strategic Plan 2020–2025 [55]
	Canada	Digital Literacy Exchange Program (DLEP) [56] Digital Inclusion Principles supporting older adults [57]
South America	Argentina	Digital Inclusion for Older People (Buenos Aires) [58]
	Brazil	Estratégia de Saúde Digital para Brasil 2020–2028 (ESD28) [59]
	Uruguay	Plan Ibirapitá – Digital Inclusion for Older Adults [60]
Ocenia	Australia	Australia's National Digital Health Strategy [61] Digital Health Blueprint & Action Plan 2023–2033 [62] Aged Care Data and Digital Strategy [63] Digital Literacy Program for Older Australians [64]
	New Zealand	Healthy Ageing Strategy 2016 [65] Better Later Life – He Oranga Kaumātua [66] Digital Inclusion Blueprint [67]
Africa	African Union	Digital Transformation Strategy for Africa 2020–2030 [68]
	South Africa	National Digital Health Strategy 2019–2024 [69]
	Rwanda	National Digital Health Strategy (Rwanda) [70]
	Nigeria	Public-private partnerships for older [71]
Global	UN	Global Strategy on Digital Health 2020–2025 [3], [33] UN Decade of Healthy Ageing 2021–2030 [72] Ageing in a Digital World – from Vulnerable to Valuable (2021) [73], [74], [75] mAgeing [76][11]

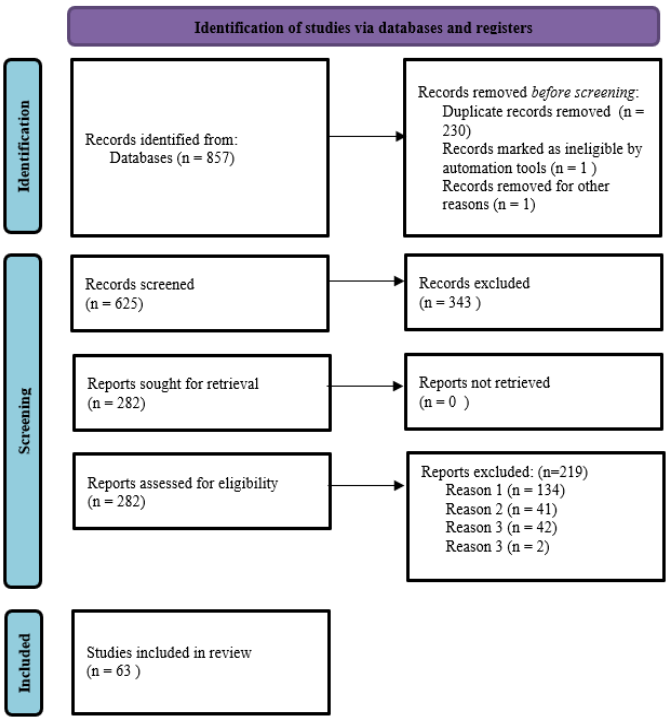


Fig. 1. The PRISMA-ScR process.

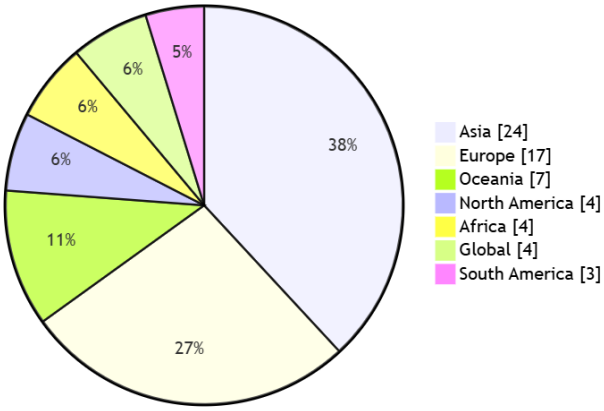


Fig. 2. Distribution of policies by region.

A. Overview of Digital Ageing Policies

The included documents represent a broad spectrum of initiatives addressing ageing populations through digital transformation. These initiatives include national digital policies, international frameworks, AI-driven healthcare initiatives, smart city ageing programmes, long-term care digital strategies, and national action plans.

The reviewed policies span multiple geographical regions, including North America, Europe, Asia, Oceania, and international organisations such as the United Nations (UN), World Health Organisation (WHO), OECD, and the European Union (EU).

Consistent with the objective of a scoping review, the included documents were analysed to identify:

- Policy priorities for ageing societies
- Digital technologies incorporated within policies
- Target domains addressed by national strategies
- Implementation mechanisms and governance approaches
- Gaps and limitations within existing frameworks

B. Geographic Distribution of Policies

The mapping of global policy initiatives indicates that digital ageing strategies are predominantly concentrated in highly developed economies where demographic pressures intersect with advanced technological infrastructures.

Within Europe, policy frameworks such as the European Union Ageing Strategy, Digital Single Market initiatives, Active and Healthy Ageing programmes, and the Ambient Assisted Living (AAL) programme demonstrate long-standing efforts to integrate digital technologies within ageing welfare systems. Similarly, advanced Asian economies have implemented comprehensive digital ageing strategies. Japan's Society 5.0, South Korea's Digital Inclusion Act, Singapore's Smart Nation initiative, and Thailand's Digital Thailand strategy illustrate the region's strong emphasis on utilising emerging technologies to enhance healthcare services, independent living, and social participation for older adults.

In North America, the United States and Canada have institutionalised digital approaches within health and social care systems through initiatives such as the Federal Health IT Strategic Plan, digital literacy programmes, virtual care, and older-adult inclusion initiatives.

In contrast, many developing countries demonstrate fragmented or early-stage digital ageing policies, often limited to pilot projects, donor-supported initiatives, or local programmes without comprehensive national frameworks. This disparity highlights a persistent global digital divide where countries with advanced digital infrastructures and stronger governance capacities are better positioned to implement comprehensive digital ageing strategies.

C. Focus Areas of Digital Ageing Policies

Analysis of the included documents revealed five major thematic clusters that characterize contemporary digital ageing strategies globally. These thematic areas reflect the multidimensional nature of ageing challenges and highlight how governments increasingly rely on digital infrastructures and Industry 4.0 technologies to support health, social participation, mobility, protection, and economic inclusion for older adults.

1) Healthcare digitalisation and smart health: Healthcare digitalisation emerged as the most dominant focus area across global ageing policies. Governments increasingly prioritise telemedicine platforms and remote consultation systems to improve healthcare accessibility for older adults, particularly those with mobility limitations or chronic illnesses.

Artificial intelligence (AI) technologies are frequently integrated into healthcare systems to support predictive analytics, early disease detection, and personalised treatment planning. In addition, IoT-based monitoring systems, including

fall detection sensors and wearable health monitoring devices, enable continuous real-time health monitoring within home environments.

In technologically advanced countries such as Japan, South Korea, and Singapore, robotics and automation are increasingly incorporated into eldercare services, supporting daily living activities and alleviating labour shortages within the caregiving workforce.

2) Social inclusion and active ageing: Another key policy focus involves enhancing social participation and digital inclusion among older adults. Many national strategies include digital literacy programmes, simplified user interfaces designed for senior citizens, and digital community platforms that facilitate civic engagement and social interaction.

Intergenerational digital hubs are increasingly established to support collaborative learning and social activities between younger and older generations. These initiatives highlight the recognition that ageing is not solely a biomedical issue but also a social phenomenon where digital connectivity plays a crucial role in maintaining psychological well-being and community participation.

3) Smart environments and age-friendly cities: Digital ageing policies increasingly integrate smart city technologies to support independent living among older adults. These initiatives include the development of smart homes equipped with automated systems, environmental sensors, adaptive lighting, and remote monitoring technologies.

Universal design principles are widely adopted to ensure accessibility across public infrastructure, housing environments, and transportation systems. Smart mobility solutions such as real-time navigation systems, barrier-free public transport, and autonomous transport technologies are also being explored to enhance mobility and independence among older adults.

Cities including Tokyo, Seoul, and Toronto have implemented integrated digital urban strategies that combine housing digitalisation, smart healthcare services, and community-based support systems.

4) Governance, protection, and data regulation: The fourth thematic domain centres on governance innovations and regulatory frameworks designed to safeguard older adults within digital ecosystems. Many policies articulate robust data protection mechanisms addressing privacy concerns, especially considering older adults' increased vulnerability to cyber risks. Governments are also developing AI ethics guidelines that regulate algorithmic decision-making in healthcare, welfare assessments, and surveillance technologies.

Digital accessibility policies, grounded in rights-based approaches, aim to ensure equitable access to digital services irrespective of age, literacy level, or socioeconomic conditions (ageism). Additionally, several countries have introduced funding mechanisms and innovation accelerators to stimulate research, commercialisation, and ecosystem development around ageing technologies.

5) Work, economics, and productivity: The final thematic focuses on the economic dimensions of ageing, particularly within the EU and OECD contexts where ageing-related labour shortages are becoming increasingly prominent. Policies in this area highlight the importance of upskilling older workers through digital learning platforms, promoting lifelong learning, and supporting flexible employment ecosystems that leverage remote work, gig platforms, and age-inclusive digital industries.

Some countries also explore automation compensation mechanisms and digital tools for labour-market reintegration, recognising that sustained economic participation among older adults is essential for national productivity and long-term economic resilience. These initiatives signal a growing awareness that digital transformation must be closely aligned with workforce longevity and economic sustainability in ageing societies.

D. Challenges and Identified Gaps

Despite the growing number of digital ageing initiatives worldwide, several systemic gaps remain evident. One major challenge involves fragmented governance structures, where responsibilities for ageing and digitalisation are distributed across multiple agencies or ministries, often leading to inconsistent policy implementation and limited cross-sector coordination.

Another recurring issue is the low level of digital literacy among older adults, particularly in rural or lower-income regions, which restricts the effectiveness of digital interventions. Furthermore, digital infrastructures in many rural areas remain underdeveloped, limiting access to telehealth services and digital public platforms.

Ethical concerns related to AI governance and algorithmic decision-making also remain insufficiently addressed within many national policies, particularly regarding issues of data privacy, algorithmic bias, and digital exploitation of vulnerable populations.

Additionally, ageing-related data are frequently fragmented across different government systems, preventing integrated data analytics and evidence-based policymaking. These structural limitations highlight the need for more coordinated and inclusive digital transformation strategies for ageing societies.

V. Discussion

A. Interpretation of Key Findings

The findings of this scoping review reveal that while digital strategies for ageing populations are expanding globally, the landscape remains highly decentralised, unevenly implemented, and shaped by cultural and socioeconomic contexts. The most advanced ageing-nation digital ecosystems, particularly Japan, South Korea, Singapore, and selected member states of the European Union, exhibit strong integration of real-time health monitoring, predictive analytics, and population health surveillance, supported by well-developed digital infrastructures. These countries have also invested heavily in integrated digital public services, smart urban ageing environments, and robust regulatory and funding governance

structures. Such systems reflect long-term institutional commitments to embedding digitalisation within broader national ageing strategies.

Despite these advancements, the review highlights that there is no universal or standardised digital transformation framework applicable across ageing nations. Policy responses remain highly context-dependent, shaped by demographic characteristics, economic capabilities, political structures, and cultural norms. This heterogeneity reinforces earlier arguments that digital ageing policies must be locally adapted rather than globally prescribed.

The predominance of digital health strategy/infrastructure entries is consistent with the global framing that positions digital health as a system-strengthening tool. WHO's global strategy provides a strong rationale for why countries prioritise governance, interoperability and standards. Without these elements, ageing-focused digital services become isolated "apps" rather than system capabilities. This pattern is visible in national strategies that emphasise coherent data-sharing and trustworthy networks (e.g., Denmark) and national platforms enabling citizen and provider access to records (e.g., Finland's Kanta services) [42].

The dataset also shows that digital inclusion & skills policies are frequently programme-based and community-facing, like training, devices, digital identity onboarding, and safer internet use. Singapore's Smart Nation trajectory illustrates how digital inclusion can be formalised as part of national digital transformation, including explicit senior training achievements and a technology portfolio that includes AI and IoT [23].

Australia's Be Connected similarly frames older-adult digital skills and online safety as an ongoing public programme with a multi-year extension, signalling that digital inclusion is treated as a continuing public good rather than a one-off campaign.

B. Cross Country Comparison

A useful cross-country distinction is "infrastructure-first" versus "inclusion-first". Many European countries are "infrastructure-first", which focus on interoperability, platform services, and regulatory governance. Several Asian entries are "transformation-first", which broad digital society programmes (Smart Nation; Society 5.0; Digital India; Indonesia's digital vision) establish a national narrative for technology-enabled development, and ageing-specific measures are nested within that narrative through care models or inclusion programmes.

Oceania is notable for combining both with Australia pairs a digital health blueprint with an aged care data and digital strategy, while also sustaining older-adult digital inclusion programs. New Zealand similarly separates "healthy ageing" and "digital inclusion" into distinct strategies with their own governance pathways, like the Healthy Ageing Strategy, Better Later Life and Digital Inclusion Blueprint, which is analytically valuable because it allows evaluation questions to be matched to the most appropriate policy instrument[77].

In South America, the strongest "evaluation-ready" signal is where programmes are concrete and bounded, like Buenos Aires legal codification of a senior digital inclusion programme and Brazil's digital health strategy with subsequent monitoring. In Africa, the continental African Union (AU) strategy provides enabling pillars like infrastructure, skills, and enabling regulation, which critically has a monitoring, evaluation, and learning (MEL) framework published for implementation learning, which offers a template for countries that lack national evaluation capacity.

C. Best Practices and Transferable Policy Mechanisms

A first transferable mechanism is indicatorisation, turning strategy into measurable commitments. Sweden's indicator-based framework for the follow-up of Vision eHealth 2025 explicitly formalises indicators and is used to structure monitoring outputs, making it a high-quality exemplar for evaluation-driven governance[44]. Similarly, the US Federal Health IT Strategic Plan frames implementation as measurable progress and connects strategy to reporting mechanisms, strengthening accountability for interoperability and access goals[55]. Brazil's ESD28 is relevant because it pairs a strategy document with subsequent monitoring and evaluation reporting, signalling an intention to track progress beyond rhetoric[59].

A second mechanism is evidence-gated adoption. Germany's DiGA fast-track approach institutionalises evaluation by requiring a formal assessment pathway for reimbursable digital health applications and anchoring adoption in a defined evidence concept. This is not directly transferable as a "copy-paste" reimbursement model, but it is transferable as a policy design pattern, with time-boxed evaluation plus conditional adoption to prevent uncontrolled diffusion of unvalidated tools for older adults.

A third mechanism is legal governance for AI and digital services. The EU AI Act operationalises a risk-based approach with obligations that directly map to evaluation and monitoring (dataset quality, traceability, cybersecurity, human oversight). For eldercare contexts where surveillance-like tools, prediction systems, or decision support may be deployed, this provides a policy vocabulary to argue for minimum auditability and safety standards.

A fourth mechanism is community-based inclusion delivery. Singapore's senior digital inclusion programme and Australia's Be Connected show a scalable approach that leverages community partners, formalises curricula and support channels, and treats inclusion as long-term capability-building. Uruguay's Ibirapitá further demonstrates the value of framing inclusion as a rights-based, participatory policy rather than merely "training sessions", which supports sustained engagement.

D. Implications for Malaysia

The findings of this scoping review have important implications for Malaysia as it prepares to become an aged nation. Malaysia has already introduced several ageing-related initiatives, including the National Policy for Older Persons, the National Plan of Action for Older Persons, Pusat Aktiviti Warga Emas (PAWE), Unit Penyayang Warga Emas (UPWE), and welfare-based community programmes. However, these initiatives remain largely social, welfare, and community-development oriented. Therefore, Malaysia's next strategic priority should be to digitally strengthen existing ageing

structures rather than develop a separate digital ageing ecosystem.

PAWE is particularly relevant because it already functions as a community platform for older persons. It can be repositioned as a local digital ageing hub that supports digital literacy, telehealth access, e-government assistance, scam-awareness education, intergenerational digital learning, and community-based social participation. Through this approach, PAWE can become an implementation channel for Malaysia's digital ageing agenda, especially among rural, low-income, and digitally excluded older adults.

At the national level, Malaysia requires a coordinated inter-ministerial governance structure for digital healthy ageing involving welfare, health, digital transformation, local government, housing, and urban planning agencies. Such coordination is consistent with the MADANI aspiration, particularly its emphasis on social protection, inclusiveness, sustainability, compassion, and good governance. In this context, digital transformation for ageing should not be viewed merely as a technological agenda, but as part of a broader national effort to build a caring, equitable, and resilient society.

The integration of AI, Big Data, and IoT into healthcare, social protection, and urban systems should be prioritised to improve predictive planning, service delivery, resource allocation, and early identification of older persons requiring support. Malaysia should also strengthen data governance to ensure privacy, consent, interoperability, cybersecurity, and human oversight in all digital ageing initiatives. This is especially important for technologies such as remote monitoring, assistive robotics, predictive analytics, and AI-enabled health or welfare services.

Collectively, these implications reinforce the strategic importance of the proposed MalaysiaAgeNet framework as a national digital transformation blueprint. By integrating global best practices with existing Malaysian platforms such as PAWE and UPWE, and aligning them with the values of MADANI and the Sustainable Development Goals, MalaysiaAgeNet can provide a practical and context-sensitive pathway for building an inclusive, ethical, and digitally enabled ageing society.

VI. Limitations

This scoping review possesses several methodological limitations that should be acknowledged. First, the scope and depth of available policies varied significantly across countries, with some providing comprehensive national strategies while others offered only high-level or fragmented policy documents. This variability may have influenced the comparative richness of thematic mapping.

Second, the review may not have captured relevant grey literature or unpublished policy documents not indexed in major academic databases, particularly from low and middle-income nations. Third, the rapid pace of technological evolution means that some of the identified policies, especially those published several years ago, may no longer reflect contemporary technological capabilities or policy directions.

Fourth, consistent with PRISMA-ScR guidance, this review did not assess the quality or effectiveness of policy implementation. Therefore, the presence of a policy does not necessarily reflect its success or impact. Finally, the availability of translated government documents was limited, particularly from non-English-speaking countries, which may have resulted in the underrepresentation of certain regions. Despite these limitations, the review provides a robust and timely mapping of global digital ageing strategies and aligns with the intended purpose of scoping reviews: to synthesise broad evidence, identify conceptual patterns, and illuminate gaps in existing knowledge.

VII. Conclusion

This scoping review provides a comprehensive mapping of global digital policies and transformation strategies addressing ageing societies. The findings indicate that digital transformation is increasingly recognised as a key policy mechanism for managing demographic ageing. While several countries have developed advanced digital ecosystems integrating healthcare digitalisation, smart living environments, and regulatory frameworks, many developing countries remain at early stages of policy development.

Persistent challenges include fragmented governance structures, insufficient digital literacy among older populations, limited infrastructure in rural areas, and inadequate attention to ethical and regulatory issues surrounding digital technologies.

For Malaysia, these findings highlight the urgent need to develop an integrated national framework for digital ageing transformation. Establishing coordinated policies and strengthening digital infrastructures will be essential to ensuring sustainable healthcare systems, social inclusion, and economic resilience in the context of an ageing population.

Acknowledgment

This research was funded by the Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Grant Scheme-Early Career Researcher (FRGS-EC) [Grant Number: FRGS-EC/1/2024/SSI11/UPSI/02/2]. The funder had no role in the study design, data collection, analysis, interpretation of results, or manuscript preparation. The authors would like to acknowledge the support of SIG CultureTech, Universiti Pendidikan Sultan Idris, for its academic support and encouragement in completing this research.

References

[1] N. Azhar and A. Omar, "Aging Workforce in Malaysia: Navigating Challenges and Shaping Policies," Quantum Journal of Social Sciences and Humanities, vol. 6, no. 2, pp. 226–236, Apr. 2025, doi: 10.55197/qjssh.v6i2.615.

[2] World Health Organization, "UN Decade of Healthy Ageing: Plan of Action," 2023.

[3] World Health Organization, "Global Strategy on Digital Health 2020-2025," World Health Organization, 2021.

[4] H. Valokivi, S. Carlo, E. Kvist, and M. Outila, "Digital ageing in Europe: a comparative analysis of Italian, Finnish and Swedish national policies on eHealth," Ageing Soc., vol. 43, no. 4, pp. 835–856, Apr. 2023, doi: 10.1017/S0144686X21000945.

[5] S. Harper, "Economic and social implications of aging societies," Oct. 31, 2014, American Association for the Advancement of Science. doi: 10.1126/science.1254405.

[6] D. Wachs, G. Demarco, and F. Stewart, "Lessons for aging countries: Navigating the challenges of a growing elderly population," World Bank

Blogs. Accessed: May 10, 2025. [Online]. Available: https://blogs.worldbank.org/en/investinpeople/Lessons-for-aging-countries

[7] D. E. Bloom and D. L. Luca, "The Global Demography of Aging: Facts, Explanations, Future," 2016.

[8] S. García-Aguña, D. Fernández-Roses, L. P. Gallardo-Peralta, and E. Sánchez-Moreno, "Social Relations and Digital Inclusion: Barriers, Facilitators and Impact on the Interactions Among Older Adults in Spain," J. Evid. Based. Soc. Work, vol. 22, no. 6, pp. 836–853, Nov. 2025, doi: 10.1080/26408066.2025.2521518.

[9] T. Greenhalgh et al., "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies," J. Med. Internet Res., vol. 19, no. 11, p. e367, Nov. 2017, doi: 10.2196/jmir.8775.

[10] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," International Journal of Social Research Methodology: Theory and Practice, vol. 8, no. 1, pp. 19–32, Feb. 2005, doi: 10.1080/1364557032000119616.

[11] R. A. Ramsawak, D. K. Behera, and S. K. Nayak, "Digital Health Solutions and Its Usage by Ageing Populations in Asia: A Bibliometric Analysis and Review," Oct. 01, 2025, SAGE Publications Inc. doi: 10.1177/21582440251396495.

[12] B. Ma et al., "Artificial intelligence in elderly healthcare: A scoping review," Jan. 01, 2023, Elsevier Ireland Ltd. doi: 10.1016/j.arr.2022.101808.

[13] D. Levac, H. Colquhoun, and K. K. O'brien, "Scoping studies: advancing the methodology," 2010. [Online]. Available: http://www.cihr-irsc.ca

[14] A. C. Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," Oct. 02, 2018, American College of Physicians. doi: 10.7326/M18-0850.

[15] C. L. Sheppard, C. Kwon, M. Yau, J. Rios, A. Austen, and S. L. Hitzig, "Aging in Place in Social Housing: A Scoping Review of Social Housing for Older Adults," Canadian Journal on Aging, vol. 42, no. 1, pp. 69–79, Mar. 2023, doi: 10.1017/S0714980822000125.

[16] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," BMC Med. Res. Methodol., vol. 18, no. 1, Nov. 2018, doi: 10.1186/s12874-018-0611-x.

[17] T. and I. C. O. G. of J. Council for Science, "[Tentative Translation] Report on The 5 th Science and Technology Basic Plan," 2015.

[18] C. Narvaez Rojas, G. A. Alomia Peñafiel, D. F. Loaiza Buitrago, and C. A. Tavera Romero, "Society 5.0: A Japanese concept for a superintelligent society," Jun. 02, 2021, MDPI AG. doi: 10.3390/su13126567.

[19] Mayumi Fukuyama, "Society 5.0 : Aiming for a New Human Centered Society," Japan Spotlight, 2018, [Online]. Available: http://www8.cao.go.jp/cstp/

[20] Cabinet Secretariat of Japan, "Japan Revitalization Strategy 2016," 2013.

[21] M. Ishihara, S. Matsunaga, R. Islam, O. Shibata, and U. Il Chung, "A policy overview of Japan's progress on dementia care in a super-aged society and future challenges," 2024, National Center for Global Health and Medicine. doi: 10.35772/ghm.2023.01047.

[22] O. Masaaki, "Topics: Recent topics in public health in Japan 2024 Community-based inclusive society and integrated care in Japan: Concepts and challenges for practice," J. Natl. Inst. Public Health, vol. 73, no. 1, pp. 32–41, 2024.

[23] S. L. Foo and G. Pan, "Singapore's vision of a smart nation," 2016. [Online]. Available: https://ink.library.smu.edu.sg/soa_research

[24] N. H. W. Ngiam et al., "Building Digital Literacy in Older Adults of Low Socioeconomic Status in Singapore (Project Wire Up): Nonrandomized Controlled Trial," J. Med. Internet Res., vol. 24, no. 12, p. e40341, Dec. 2022, doi: 10.2196/40341.

[25] Gregoire Virat, "Smart Care at Home: Scaling Singapore AI Home Healthcare for Aging Populations," Eurogroup Consulting.

[26] Z. Jiang and W. Jiang, "Health Education in the Healthy China Initiative 2019−2030," China CDC Wkly., vol. 3, no. 4, pp. 78–80, Jan. 2021, doi: 10.46234/ccdcw2021.018.

[27] W. Lee and Y. Kim, "A Study of the Legal Framework for Digital Inclusion in South Korea," Korea Policy Journal, vol. 23, pp. 36–52, Dec. 2023, doi: 10.62353/kopoj.2023.23.36.

[28] Seung Hyun Kim, "Future Policy Directions for Planning of National Responsibility for Dementia Care," 2017.

[29] OECD Observatory of Public Sector Innovation, "Enhancing Seoul's Age Readiness, the Comprehensive Plan for 50+ Assistance," 2021.

[30] Marketing & Communications, "Thailand Digital Health Transformation Gains Speed."

[31] S. Konglertmongkol, "Mainstreaming Digital Literacy into Ageing Policies: the Way Forward to Empower Older Persons in Thailand," 2024. [Online]. Available: www.dop.go.th

[32] K. Robru, P. Setthasuravich, A. Pukdeewut, and S. Wetchakama, "Internet Use for Health-Related Purposes among Older People in Thailand: An Analysis of Nationwide Cross-Sectional Data," Informatics, vol. 11, no. 3, Sep. 2024, doi: 10.3390/informatics11030055.

[33] S. Jantavongso, "Toward global digital literate citizens: A case of Thailand's aging generation," Electronic Journal of Information Systems in Developing Countries, vol. 88, no. 2, Mar. 2022, doi: 10.1002/isd2.12207.

[34] Ministry of Public Health Qatar, "The National Health Strategy 2018-2022 (Qatar)," 2018.

[35] G. Albahri and H. Alothman, "Elderly in the UAE," Al-Adab Journal, no. 152, pp. 129–164, Mar. 2025, doi: 10.31973/g6596456.

[36] K. Maurya, "From Pixels to Policies: Understanding the Impact of Digital Governance Initiatives in India," 2024.

[37] M. A. Alifiana, M. M. Hanafi, and B. Setiyono, "Digital Entrepreneurship and Financial Performance: Innovation Adoption in Women-Led Indonesian MSMEs," 2025, pp. 273–289. doi: 10.1007/978-981-95-1148-8_18.

[38] European Commission, "Green Paper on Ageing Fostering solidarity and responsability between generations," 2021.

[39] European Union, "European Innovation Partnership on Active and Healthy Ageing," https://digital-strategy.ec.europa.eu/en/policies/eip-aha.

[40] G. Aumayr, "From Ambient Assisted Living to Active and Assisted Living: A Practical Perspective on Experiences and Approaches," 2016, pp. 3–13. doi: 10.1007/978-3-319-39904-1_1.

[41] J. Sattar, "Navigating the Grey Area: Legal Frameworks for Digital Health Monitoring & Use of AI for Elderly Patients in the Nordics," International Journal of Technology and Emerging Sciences, 2024, [Online]. Available: www.mapscipub.com

[42] V. Jormanainen, "Funding Nationwide HIE and HIS: Case Kanta Services 2003–2024 in Finland," 2025. doi: 10.3233/SHTI250134.

[43] Swedish Government Official Reports, Effektiv vård : slutbetänkande. Wolters Kluwer, 2016.

[44] SALAR, "Vision for eHealth 2025," 2025.

[45] Government of Italy, "Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 6: Salute.," 2025.

[46] Secrétariat Général Pour L'Investissement Et Ministère De La Transition Écologique, "L'évaluation socioéconomique des effets de santé des projets d'investissement public," 2025. [Online]. Available: www.strategie.gouv.fr

[47] S. Gerke, A. D. Stern, and T. Minssen, "Germany's digital health reforms in the Covid-19 era: lessons and opportunities for other countries," Dec. 01, 2020, Nature Research. doi: 10.1038/s41746-020-0306-7.

[48] E. Burk, J. Hornung, and C. Schoppe, "The Digital Act: a boost for digitalisation in German healthcare." Accessed: Feb. 20, 2026. [Online]. Available: https://www.gleisslutz.com/en/know-how/digital-act-boost-digitalisation-german-healthcare

[49] M. Romanyuk, M. Sukharnikova, and N. Chekmareva, "Trends of the digital economy development in Russia," in IOP Conference Series: Earth and Environmental Science, IOP Publishing Ltd, 2021. doi: 10.1088/1755-1315/650/1/012017.

[50] J. Rasmussen, "eHealth Governance - Country Report: Denmark," 2021.

[51] Hon Greg Clark, "Government announces £300 million for landmark ageing society grand challenge," Gov.uk, 2018.

[52] S. Thomas, "The NHS England Long Term Plan implementation framework," British Journal of Neuroscience Nursing, vol. 15, no. 4, pp. 184–186, Aug. 2019, doi: 10.12968/bjnn.2019.15.4.184.

[53] C. Artigas Brugal, "España digital: la hoja de ruta para impulsar un país en clave de futuro," Economia Digital, pp. 352–356, 2023.

[54] E. L. Burd, S. P. Smith, and S. Reisman, "Exploring Business Models for MOOCs in Higher Education," Innov. High. Educ., vol. 40, no. 1, pp. 37–49, 2014, doi: 10.1007/s10755-014-9297-0.

[55] National Coordinator for Health Information Technology, "2020-2025 Federal Health IT Strategic Plan," 2021.

[56] Agewell, "Digital Literacy Interventions for Older Adults in Canada," 2021.

[57] B. B. Neves and G. Mead, "Digital Technology and Older People: Towards a Sociological Approach to Technology Adoption in Later Life," Sociology, vol. 55, no. 5, pp. 888–905, Oct. 2021, doi: 10.1177/0038038520975587.

[58] WHO, "+Simple, digital inclusion for older people," WHO, 2018.

[59] Ministério da Saúde, "Estratégia de Saúde Digital para o Brasil 2020-2028," 2020.

[60] Q. Gloria, "Nearly 4 years in, Uruguay's Plan Ibirapitá promotes digital inclusion for the country's senior citizens," Medium, 2019.

[61] Australia Digital Health Agency, "Australia's National Digital Health Strategy," 2021.

[62] Commonwealth of Australia, "Digital Health Blueprint 2023-2033," 2023, [Online]. Available: http://www.dpmc.gov.au/government/commonwealth-coat-arms

[63] Department of Health and Aged Care, "National Aged Care Data Strategy Acknowledgement of Country," 2023.

[64] A. McCosker et al., "Improving the digital inclusion of older Australians: The social impact of Be Connected," 2020. doi: 10.25916/5ed5b6e204a95.

[65] Associate Minister of Health, Healthy Ageing Strategy. Ministry of Health, 2016.

[66] A. Akinwande, "Successful ageing and trajectories of health among New Zealand older adults," 2022.

[67] Department of Internal Affairs Te Tari Taiwhenua, The Digital Inclusion Blueprint. Department of Internal Affairs Te Tari Taiwhenua, 2019.

[68] African Union, "The Digital Transformation Strategy For Africa (2020-2030)," 2020. [Online]. Available: www.au.int

[69] National Dept Health South Africa, "National Digital Health Strategy for South Africa 2019 - 2024," 2020.

[70] G. Pearce, "The Benefits of National Digital Health in a Short Story," ISACA. Accessed: Nov. 22, 2025. [Online]. Available: https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/the-benefits-of-national-digital-health-in-a-short-story?utm_source=chatgpt.com

[71] J. I. ANI and K. Batisai, "Promoting digital inclusion through public-private partnerships for older adults in Nigeria: A review," Aging Health Res., vol. 4, no. 4, Dec. 2024, doi: 10.1016/j.ahr.2024.100211.

[72] United Nations, "UN Decade of Healthy Ageing: Plan of Action," 2021.

[73] International Telecommunication Union, "Ageing in a digital world – from vulnerable to valuable," 2021.

[74] A. J. Scott, "The longevity economy," Dec. 01, 2021, Elsevier Ltd. doi: 10.1016/S2666-7568(21)00250-6.

[75] Oxford Economics, "The Longevity Economy Generating economic growth and new opportunities for business," 2022.

[76] World Health Organization and International Telecommunication Union, "BE HE@LTHY BE MOBILE : A Handbook on how to implement mAgeing," 2018. [Online]. Available: http://apps.who.int/bookorders.

[77] Ministry of Social Development New Zealand, "Better Later Life 2019 - 2034," 2019.

A Mid-Level Feature Fusion Framework Integrating PHLF and VGG16 for Robust Batik Pattern Detection

Jani Kusanti[1]*, Edi Noersasongko[2], Purwanto[3], Moch Arief Soeleman[4]
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia [1, 2, 3, 4]
Faculty of Electrical and Information Engineering, Universitas Surakarta, Surakarta, Indonesia [1]

***Abstract*—Javanese batik is characterized by distinctive motif patterns. The rapid evolution of designs through motif combination has introduced increased complexity, posing challenges for the community. People are increasingly less familiar with traditional Javanese original batik patterns. Many studies using CNN have been conducted to recognize batik patterns. However, it is important to improve object detection performance by strengthening explicit local features. To answer this question, our study aims to improve the detection of batik patterns that have local texture patterns and complex orientations that are in harmony with batik geometry with an integrated fusion scheme. To improve detection, an integrated fusion schema model was developed using the PHLF hybrid framework as a geometrically oriented local feature with VGG16 as a deep feature extractor for object detection. According to research, although VGG16 is reliable on large benchmarks such as ImageNet, VGG16 is less reliable for subtle intra-motif variations, which can suppress accuracy. Evidence from the results of the study shows that in the eight-class dataset consisting of 6,400 images for training data and 1.600 images for test data, a mid-level feature fusion approach on VGG16 with integrated PHLF improves resistance to data variations such as lighting, fabric deformation, and background complexity. The experimental results showed that the model achieved an mAP value of 0.68 at IoU = 0.5 and 0.2846 at IoU = 0.7. The significant difference between mAP@0.5 and mAP@0.7 suggests that the model still has limitations in the precision of the localization of the boundary box.**

Keywords*—A Mid-level feature; detection; fusion framework; phlf; traditional javanese batik

I. Introduction

Batik, an intangible cultural heritage with spiritual value, a deep history, and philosophy [1][2]. However, the characteristics of batik that are complex, repetitive, and have a high resemblance between classes are a challenge in batik detection [3][4]. To answer the challenge, various studies have been conducted, from the use of Deep Learning models such as Convolutional Neural Network (CNN) using VGG16, ResNet, and EfficientNet architectures, proven to be effective in capturing fine texture details [3][5]. Research shows that the Transfer Learning Technique and Hybrid models of VGG16 and XGBoost can achieve up to 89.83% accuracy in recognizing traditional motifs [3][6]. However, the research that has been conducted has not answered the problem, considering that Batik has undergone such rapid development. The use of CNN architecture is the dominant method in batik analysis because of its ability to automatically learn features directly from the data used [7][8]. Other architectures, such as Inception ResNetV2, have proven their effectiveness by achieving an average accuracy of 98.19% in the classification of batik patterns [9], while the VGG16 and MobileNetV2 models also show high results in detecting complex visuals [10]. For real-time detection, the Yolov11 model enhanced with an attenuation mechanism was able to achieve an mAP (mean Average Precision) value of 0.748. Provides a balance between localization accuracy and computational efficiency [11].

Although CNN excels in extracting global semantic features, this model has limitations in representing the subtle textures that characterize batik [12][13][14][15]. In contrast, feature selection such as the Gray Level Co-occurrence Matrix (GLCM) is highly effective in describing local texture characteristics through the spatial relationships between pixels [16][17][18][19]. However, this method often faces the challenge of semantic gaps and is less adaptive to large data variations in real time [20]. To answer the challenge, several studies with a hybrid approach and feature fusion were conducted, combining models is a crucial strategy to improve the performance of automatic detection [3][20]. One of the models is the combination of VGG16 and XGBoost, which managed to achieve the highest accuracy in Batik Keraton at 89.83% compared to the CNN model [3]. The integration of features in latent space, as implemented in the Deep Convolutional Autoencoder (DCAE) model, has been shown to produce a more discriminatory, clean representation of noise with an accuracy of up to 99% [21]. Commonly used fusion strategies include concatenation at the feature level, which is considered more advantageous because it can store different information from each extraction method [20]. In addition, the use of dual-attention modules that combine spatial attention and channel attention significantly strengthens the model's focus on the most informative motif regions [12]. Hybrid-based detection models that integrate VGG16 for high-level semantic features with local texture orientation features are strongly aligned with the finding that the combination of shape and texture features results in greater accuracy (90.48%) than Single features [22][9]. The multi-task learning approach that performs simultaneous bounding box regression and classification, as applied to the YOLO series, is an effective solution to the dense and complex challenge of localizing batik objects [11].

Based on the results of the study, although the multi-task learning approach, such as the YOLO series, is considered effective, there are some fundamental weaknesses when applied to the dense and complex detection of batik motifs. This is important to underline the need for the development of a mid-level feature fusion framework that integrates PHLF's handcrafted features with VGG16's deep learning features. Therefore, the integration of PHLF to capture local geometric

orientations and VGG16 for high-level semantic features through intermediate-level fusion is a crucial solution to improve the identification accuracy as well as the precision of bounding box localization on batik objects.

II. Previous Research Study

Research shows that pure deep learning models often struggle with rotational and scale variations in dense batik motifs. In the study conducted by [23], the MU2ECS-LBP and Mulwin-LBP algorithms were developed specifically designed to be invariant to scale and rotation, as well as in the studies [24] and [25]. This study proves that the handling of geometric orientation specifically is crucial because shooting batik in the real world often involves inconsistent zoom and angle changes. Research conducted by [26] and [27], uses Geometric Invariant Moment (GIM) as a shape-based feature extractor to detect the Broken Parang motif. This supports the argument that geometric information and local shapes are formidable descriptors for recognizing the structure of a particular motif.

VGG16 is recognized in many studies as having the ability to extract very deep spatial features for intricate patterns. As in the research conducted [3], it is explained that the architecture of VGG16 is very effective in capturing the delicate textures and intricate ornaments of batik through its deep and dense convolutional structure. The VGG16 model is able to capture subtle differences in texture that traditional models often fail to identify. Likewise, in the study [28], utilizing VGG16 to study the unique characteristics of batik spatially through the convolution process proves that visual-based feature extraction is far superior to simple color-based methods. The integration between traditional features and deep learning features has been shown to mask the weaknesses of each method, as in studies conducted [29] and [24], it was found that the combination of texture features and shape features resulted in much higher accuracy (90.48%) than using only one type of feature. Similarly, in the study [16], it was emphasized that relying only on artificial neural networks can lead to the loss of important local textural information. Therefore, fusion between handcrafted texture features and deep learning models (such as ResNet or VGG) is strongly recommended to create a more robust classification system [30]. In the research conducted [20], it is hypothesized that various image features (geometry, points, parallel lines) complement each other in representing the properties of batik motifs, and the fusion of features consistently outperforms the performance of individual features.

To identify the limitations of standard detection models in handling the density of batik motifs, the research conducted by [11], explains that object detection architectures such as the standard YOLO have limitations in the local receptive field, making it difficult to model the long-range spatial relationships that define the composition of batik culture. In addition, very small and dense batik objects often exceed the detection capabilities of the baseline model, thus requiring the extraction of higher-quality features to improve the precision of localization. Research by [3], showed that combining the extraction of VGG16 in-depth features with the XGBoost classifier resulted in the highest accuracy of 89.83% on Keraton Batik, outperforming the standard CNN model. The study by [2], [28] and [31], utilized VGG16 to study the unique characteristics of batik spatially through convolution and used Random Forest to achieve 97.58% accuracy on a dataset with 50 classes. VGG16 Modified (Web-Based), as in the study [32], modified VGG16 for the classification of Demak Batik and managed to achieve 98.72% accuracy in just 5 epochs.

The problem with the research that has been conducted is that standard *deep learning* models, including CNN and detection architectures such as YOLO, tend to lose crucial micro-texture details in batik due to repeated downsampling processes [12], fine line details are often erased or become blurred, requiring additional handcrafted features that are more sensitive to local textures to retain the information [33]. Although CNN excels in global semantic features, it often struggles to cope with extreme rotation and scale variations without the support of massive augmented data or specific geometric features [22] [3]. There is a gap between the visual representation of humans of batik motifs that are rich in cultural significance and the low-level features stored in digital databases [22][20]. Single models often fail to bridge this gap, deep learning features are too abstract for the technical details of the craftsman, while traditional features are too rigid for visual variation [16].

To address these gaps, this study proposes a hybrid approach that integrates PHLF's handcrafted texture features with the deep learning features of VGG16 through a mid-level feature fusion strategy. This approach allows for a more effective interaction between texture features and semantic features in latent spaces, resulting in more discriminatory representations for batik motif detection.

III. Proposed Research Methodology

The methodology for this study is illustrated in Fig. 1.

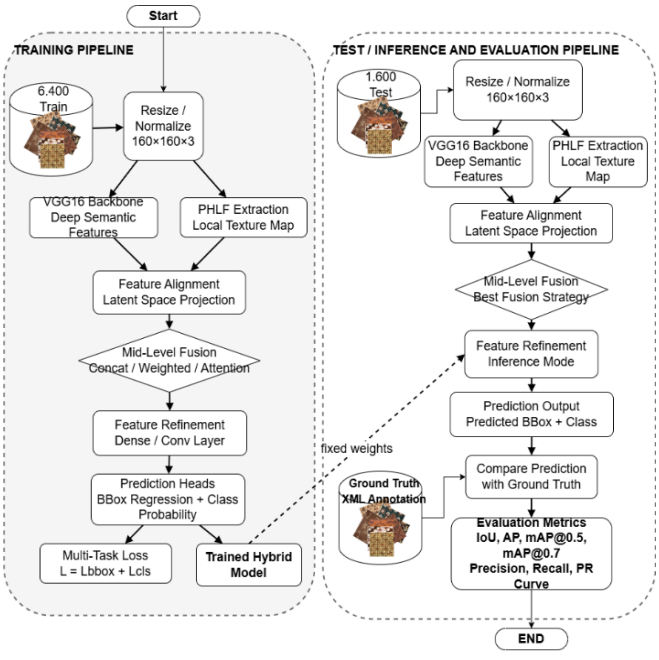


Fig. 1. Proposed method.

This study proposes a hybrid-based batik motif detection framework that integrates the handcrafted texture features of

Parallelogram Haar-Like Features (PHLF) with the deep learning features of VGG16. This approach is designed to overcome the limitations of a single method in representing the complexity of batik patterns, especially in capturing microtextures and semantic structures simultaneously.

A. Data Processing

This study uses the Batik Keraton dataset consisting of eight pattern classes, each containing 800 original images (a total of 6,400 images) for training data and 200 (a total of 1,600 images) for testing data. Fig. 2 shows the eight patterns of Batik classes used.



Fig. 2. Eight batik classes patterns used.

B. Input Processing

Given an input batik image, shown in Eq. (1):

$$I \in \mathbb{R}^{HxWx3} \tag{1}$$

The image is first normalized, shown in Eq. (2):

$$I_{norm} = \frac{I}{255} \tag{2}$$

And resized to a fixed dimension, shown in Eq. (3):

$$I_{resized} \in \mathbb{R}^{160x160x3} \tag{3}$$

A grayscale version of the image is also generated for texture feature extraction, shown in Eq. (4):

$$I_{gray} = 0.299R + 0.587G + 0.144B \tag{4}$$

Remarks:

I = original input batik image, H = height image, W = image width, 3 = number of color channels, I_{norm} = image of normalization results, $I_{resized}$ = image resized, I_{gray} = image grayscale, R, G, B = color channels red, green, blue.

C. PHLF-Based Texture Feature Extraction

The first branch focuses on extracting local texture features using Parallelogram Haar-Like Features (PHLF). This method captures directional patterns and geometric structures inherent in batik motifs.

The convolution operation is defined as shown in Eq. (5),

$$F_k(x,y) = (I_{gray} * K_k)(x,y) \tag{5}$$

where, K_k represents the PHLF kernels. The final PHLF map is computed as shown in Eq. (6),

$$PHLF(x,y) = \sum_{k=1}^{K} |F_k(x,y)| \tag{6}$$

Followed by normalization, shown in Eq. (7),

$$PHLF_{norm} = \frac{PHLF - min}{\max - min} \tag{7}$$

The resulting feature map is shown in Eq. (8).

$$X_{phlf} \in \mathbb{R}^{HxWx1} \tag{8}$$

Remarks:

x, y = pixel coordinates in the image, K_k= Colonel Philf K-K, K = total number of PHLF kernels, $F_x(x,y)$ = the characteristic response of the result of the convolution of the kernel k at position (x,y), $PHLF(x,y)$= PHLF feature value at position (x,y), $PHLF_{norm}$ = Normalized PHLF, min = Minimum values on feature maps, max = Maximum value on the feature map, X_{phlf} = feature map of PHLF extraction results.

D. Deep Feature Extraction Using VGG16

The equation for showing the convolution process at the CNN layer is shown in Eq. (9).

$$F^{(l)} = \sigma(W^{(l)} * F^{(l-1)} + b^{(l)}) \tag{9}$$

Remarks:

$F^{(l)}$ = feature map at layer l. $F^{(l-1)}$ = feature map from the previous layer, $W^{(l)}$ = convolutional weights at layer l. $b^{(l)}$= bias term at layer l . σ – activation function (e.g., ReLU, sigmoid).

The equation used to convert the PHLF and VGG16 features to compatible latent dimensions is shown in Eq. (10)-(11).

$$Z_{phlf} = f_{phlf}(X_{phlf}) \tag{10}$$

$$Z_{cnn} = f_{cnn}(X_{cnn}) \tag{11}$$

Remarks:

Z_{phlf} = Latent representation of PHLF features, Z_{cnn} = Latent representation of CNN features, $f_{phlf}(.)$ = Projection function for PHLF features, $f_{cnn}(.)$= Projection function for CNN features.

E. Concatenation Fusion

Combines features by stacking them, shown in Eq. (12).

$$F = [Z_{cnn}, Z_{phlf}) \tag{12}$$

Balances contributions from CNN and PHLF features, shown in Eq. (13).

$$F = \alpha Z_{cnn} + (1-\alpha) Z_{phlf} \tag{13}$$

Dynamically weight feature importance using attention, shown in Eq. (14)-(15).

$$A == \alpha(W_\alpha . Z_{cnn}) \tag{14}$$

$$F = A \odot Z_{cnn} + (a - A) \odot Z_{phlf} \tag{15}$$

Enhances feature representation before prediction, as shown in Eq. (16).

$$F_{ref} = \sigma(W_f . F + b_f) \tag{16}$$

F. Evaluation

Bounding Box Loss measures localization error using L1 loss, shown in the following Eq. (17).

$$L_{bbox} = ||B - \hat{B}||_1 \quad (17)$$

Intersection over Union (IoU), shown in the following Eq. (18).

$$IoU = \frac{|B_{pred} \cap B_{gt}|}{|B_{pred} \cup B_{gt}|} \quad (18)$$

Average Precision (AP), shown in the following Eq. (19).

$$AP = \int_0^1 Precision(Recall) dRecall \quad (19)$$

Mean Average Precision (mAP), shown in the following Eq. (20).

$$mAP = \frac{1}{C} \sum_{C=1}^{C} AP_C \quad (20)$$

IV. Experimental Results and Analysis

The evaluation of the performance of the VGG16 technique with integrated PHLF was verified based on a dataset of 8 batik patterns of Kawung, Parang, Satriomanah, Sidoasih, Sidomukti, Tambal, Truntum, and Wahyutumurun. The batik data set consists of 8,000 images of batik patterns, each consisting of 1,000 data points. The images are categorized according to batik patterns. Evaluation was also carried out on 1,600 data points before augmentation.

TABLE I. Train Results of 8 Batik Classes

Method	Task	Accuracy / mAP	Limitation
VGG16 + XGBoost	Classification	89.83%	No localization
VGG16 + RF	Classification	97.58%	Not detection-based
Modified VGG16	Classification	98.72%	No bounding box
YOLO	Detection	0.174 mAP	Weak for dense patterns
CNN only	Classification	–	Loss of texture detail
PHLF + VGG16	Detection	0.208 mAP@0.5	Localization still limited

A detailed description of the test results of the augmentation is shown in Table I. Fig. 3 shows the results of Accuracy and Loss, Precision, and Recall of Model VGG16 with integrated PHLF. Fig. 3 shows that the PHLF-VGG16 Hybrid model experienced a rapid increase in accuracy during the initial and stable period around 10 years of age. Training accuracy and validation remained aligned with minimal deviations, indicating effective learning with no signs of overfitting. The loss curve showed an equally consistent decrease, with validation loss tracking training loss closely, confirming good generalization to unseen data. These patterns showed a stable and efficient convergence process, supporting the conclusion that integrating PHLF with VGG16 improved feature extraction quality and improved overall training stability.

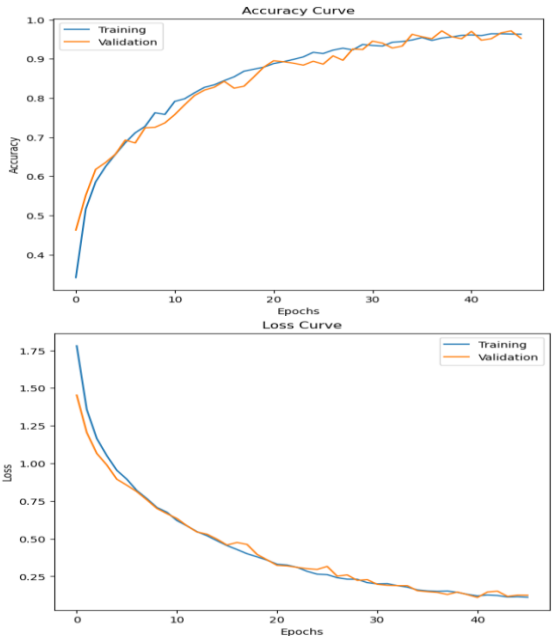


Fig. 3. Curve of accuracy and loss.

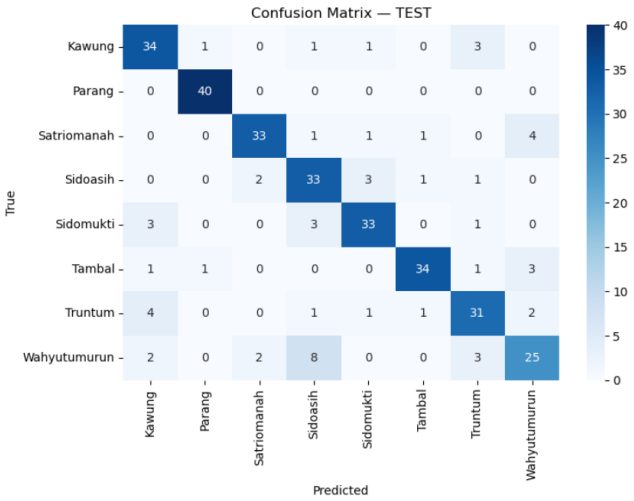


Fig. 4. Confusion matrix results test.

The Confusion Matrix shown in Fig. 4 provides strong evidence of the robustness and reliability of the VGG16 fine-tuning method using PHLF.

The sharp rise and stabilization of macro precision suggest that this method produces very few false positives, even when faced with visually similar batik motifs. This behavior suggests that the PHLF enhancement module effectively reinforces the local texture differentiation that is important for batik classification. Likewise, the convergence of F1's macro and macro graven towards unity reflects the model's ability to recognize diverse variations in motifs without systematically ignoring specific classes. This is especially important in datasets with rich intra-class variations such as Kawung, Parang, Truntum, and Tambal.

TABLE II. COMPARISON OF ACCURACY, PRECISION, RECALL, AND F1 SCORE RESULTS

Model	Acc	Precision	Recall	F1-score
VGG + XGBoost	89,83%	92%	92%	92%
VGG16 Pretrained	87,77%	88%	88%	88%
VGG16	85%	85%	85%	85%
VGG16 + PHLF	98,72%	97,6%	97,5%	97,55%

Table II shows the results, showing a consistent trend: performance improves when explicit geometry features (PHLFs) are integrated into the VGG16. In other words, the scale of the data and the feature fusion strategy are decisive, especially as the number of classes grows and the variety of motifs becomes more complex. In the initial scenario, the VGG+XGBoost approach provides the highest accuracy (89.83%) compared to the standard VGG16 ($\approx$ 85 – 87%). These results indicate that in a simpler classification space, the output of VGG features combined with powerful classifiers such as XGBoost is quite effective in separating motifs that have a pronounced visual difference.

Meanwhile, VGG16 is in the F1 range of 85%: VGG16 is capable of capturing global textures and patterns, but is still less sensitive to the diagonal structure and geometric repetition that characterize some motifs. The advantages of the VGG16 approach using PHLF clearly reach F1 = 97.55%. This spike is proof that the PHLF really does add important information: it strengthens the response to repetitive geometric elements, allowing the model to become more stable in distinguishing motifs that rely on diagonals without losing the ability of VGG16 to capture the context of texture. These findings point to two points: first, VGG16 is indeed greatly aided by a mid-level feature fusion framework that integrates PHLF and VGG16 for strong batik pattern detection; second, PHLF still provides additional advantages even though the backbone of VGG16 is already strong, resulting in more stable and consistent performance. In summary, the results of the table confirm that the enhancements in VGG16 using PHLF result in the most complete and most effective representation, especially in multi-class scenarios.

Fig. 5 presents example visualizations of batik motif detection results using the hybrid PHLF–VGG16 model, comparing ground truth (GT) and predicted outputs with evaluation metrics including TP, FP, FN, and IoU. In the left image (Kawung), the model achieved strong localization performance with mIoU(TP) $\approx$ 0.84, TP = 1, FP = 0, and FN = 0, indicating accurate detection without missed or incorrect predictions. This result suggests that the regular and repetitive structure of Kawung is easier for the model to learn and localize. In contrast, the right image (Sidoasih) shows limitations in localization precision. Although the motif was detected (FN = 0), the presence of FP = 1 and an mIoU(TP) close to zero indicates weak overlap between the predicted bounding box and the ground truth. These results suggest that the model can recognize motif presence but still struggles to precisely estimate object boundaries. The comparison highlights that detection performance is highly influenced by motif characteristics, where denser and more complex textures increase localization difficulty.

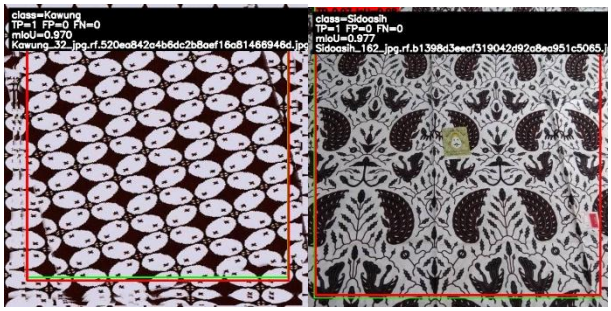


Fig. 5. Test results.

TABLE III. HASIL RATA-RATA IOU PER CLASS

Name	meanIoU	IoU@0.5
Kawung	meanIoU=0.538248	IoU@0.5=60.69%
Parang	meanIoU=0.605074	IoU@0.5=72.43%
Satriomanah	meanIoU=0.539754	IoU@0.5=60.89%
Sidoasih	meanIoU=0.512537	IoU@0.5=54.93%
Sidomukti	meanIoU=0.712856	IoU@0.5=82.22%
Tambal	meanIoU=0.546839	IoU@0.5=57.95%
Truntum	meanIoU=0.461856	IoU@0.5=47.16%
WahyuTumurun	meanIoU=0.686640	IoU@0.5=83.84%

Table III presents the batik motif detection results using mean Intersection over Union (mIoU) and IoU@0.5 to evaluate localization performance based on the overlap between predicted bounding boxes and ground truth annotations. Overall, all classes achieved IoU@0.5 above 50%, indicating satisfactory localization capability despite performance variations caused by differences in texture, object scale, and pattern complexity.

The best results were achieved by WahyuTumurun (83.84%) and Sidomukti (82.22%), demonstrating strong localization accuracy, followed by Parang (72.43%) and Kawung (60.69%). In contrast, Satriomanah, Sidoasih, Tambal, and Truntum obtained lower IoU@0.5 values (54–61%), suggesting that although objects were successfully detected, the predicted bounding boxes were not yet fully aligned with the ground truth. This limitation is likely related to texture similarity across motifs, pattern variations, small object regions, and the tendency to predict broader fabric areas than the annotated regions.

V. CONCLUSION

This study proposes a mid-level feature fusion framework that integrates handcrafted Parallelogram Haar-Like Features (PHLF) with deep semantic representations extracted by VGG16 for batik pattern detection. By combining local texture enhancement and high-level semantic learning, the framework improves feature discriminability and achieves effective detection and localization performance across all classes, with IoU@0.5 above 50% and average IoU exceeding 0.68. The best results were obtained for WahyuTumurun (83.84%) and Sidomukti (82.22%). However, performance declined at IoU = 0.7 (average IoU = 0.2846), revealing limitations in precise localization for dense and repetitive motifs due to texture similarity, scale variation, and background complexity. Overall, the proposed framework enhances texture representation and

delivers more robust, accurate, and consistent batik detection, demonstrating strong potential for real-world automated batik identification. Future work will focus on integrating Feature Pyramid Networks (FPN), attention mechanisms, and IoU aware loss functions to further improve localization accuracy for visually similar motifs.

REFERENCES

[1] V. J. Basiroen, "Batik Lasem motifs dataset: Gunung Ringgit, Kricak, Latohan, Nyuk Pitu, Seritan," 2025. doi: 10.1016/j.dib.2025.111866.

[2] D. G. T. Meranggi, "Batik Classification Using Convolutional Neural Network with Data Improvements," Int. J. Informatics Vis., vol. 6, no. 1, pp. 6–11, 2022, doi: 10.30630/joiv.6.1.716.

[3] S. Suyahman, "VGG-Based Feature Extraction for Classifying Traditional Batik Motifs Using Machine Learning Models," Preserv. Digit. Technol. Cult., 2025, doi: 10.1515/pdtc-2025-0009.

[4] A. E. Minarno, I. Soesanti, and A. Nugroho, "Batik Classification using Microstructure Co-occurrence Histogram," vol. 8, no. March, pp. 134–140, 2024.

[5] I. D. Susanti, "Yogyakarta Batik Image Classification Based on Convolutional Neural Network," Adv. Sustain. Sci. Eng. Technol., vol. 6, no. 1, 2024, doi: 10.26877/asset.v6i1.18002.

[6] M. Utami, "Improving the Transfer Learning for Batik Besurek Textile Motif Classification," Iaes Int. J. Artif. Intell., vol. 14, no. 4, pp. 3172–3181, 2025, doi: 10.11591/ijai.v14.i4.pp3172-3181.

[7] Farida, R. E. Caraka, T. W. Cenggoro, and B. Pardamean, "Batik Parang Rusak Detection Using Geometric Invariant Moment," 1st 2018 Indones. Assoc. Pattern Recognit. Int. Conf. Ina. 2018 - Proc., pp. 71–74, 2018, doi: 10.1109/INAPR.2018.8627000.

[8] C. U. Khasanah, E. Utami, and S. Raharjo, "Implementation of Data Augmentation Using Convolutional Neural Network for Batik Classification," 2020 8th Int. Conf. Cyber IT Serv. Manag. CITSM 2020, pp. 20–24, 2020, doi: 10.1109/CITSM50537.2020.9268890.

[9] B. D. Satoto, "Classification of Batik Patterns Using Inception-ResNetV2 with Data Augmentation," J. Adv. Inf. Technol., vol. 17, no. 1, pp. 42–54, 2026, doi: 10.12720/jait.17.1.42-54.

[10] E. Winarno, "Enhanced Semarang batik classification using deep learning: a comparative study of CNN architectures," Bull. Electr. Eng. Informatics, vol. 14, no. 5, pp. 3544–3557, 2025, doi: 10.11591/eei.v14i5.9347.

[11] Y. Li, "Research on batik image pattern detection based on improved YOLOv11," Npj Herit. Sci., vol. 14, no. 1, 2026, doi: 10.1038/s40494-026-02404-y.

[12] Y. Ma, "Attention-Weighted Hierarchical Decoding for Few-Shot Semantic Segmentation: A Case Study on Batik Cultural Heritage Patterns," Electron. Switz., vol. 15, no. 6, 2026, doi: 10.3390/electronics15061242.

[13] A. C. I. Ardison, M. J. R. Hutagalung, R. Chernando, and T. W. Cenggoro, "Observing Pre-Trained Convolutional Neural Network (CNN) Layers as Feature Extractor for Detecting Bias in Image Classification Data," CommIT J., vol. 16, no. 2, pp. 149–158, 2022, doi: 10.21512/commit.v16i2.8144.

[14] A. Y. Chandra, "Comparative Analysis of CNN Architectures' Performance with Vision Transformer for Batik Type Classification," 2025. doi: 10.1109/ICISIT66233.2025.11402957.

[15] S. Ariessaputra, V. H. Vidiasari, S. M. Al Sasongko, B. Darmawan, and S. Nababan, "Classification of Lombok Songket and Sasambo Batik Motifs Using the Convolution Neural Network (CNN) Algorithm," Int. J. Informatics Vis., vol. 8, no. 1, pp. 38–44, 2024, doi: 10.62527/joiv.8.1.1386.

[16] L. Elvitaria, "A Proposed Batik Automatic Classification System Based on Ensemble Deep Learning and GLCM Feature Extraction Method," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 553–561, 2024, doi: 10.14569/IJACSA.2024.0151058.

[17] Nuraedah, "Quadratic Support Vector Machine for the Bomba Traditional Textile Motif Classification," Indones. J. Electr. Eng. Comput. Sci., vol. 11, no. 3, pp. 1004–1014, 2018, doi: 10.11591/ijeecs.v11.i3.pp1004-1014.

[18] D. Wijaya, "Batik classification using KNN algorithm and GLCM features extraction," 2024. doi: 10.1051/e3sconf/202447502012.

[19] Haryanto, "BGP Model: Optimized Feature Engineering for Batik Parang Classification Using GLCM and PCA," 2025. doi: 10.1109/ICSINTESA68165.2025.11413629.

[20] I. Nurhaida, H. Wei, R. A. M. Zen, R. Manurung, and A. M. Arymurthy, "Texture fusion for batik motif retrieval system," Int. J. Electr. Comput. Eng., vol. 6, no. 6, pp. 3174–3187, 2016, doi: 10.11591/ijece.v6i6.12049.

[21] M. F. Dzulqarnain, A. Fadlil, and I. Riadi, "Improving the Accuracy of Batik Classification using Deep Convolutional Auto Encoder," vol. 13, no. 2, pp. 123–130, 2024, doi: 10.28989/compiler.v13i2.2649.

[22] A. H. Rangkuti, "A Novel Reliable Approach for Image Batik Classification That Invariant with Scale and Rotation Using MU2ECS-LBP Algorithm," 2021. doi: 10.1016/j.procs.2021.01.075.

[23] A. H. Rangkuti, "Reliable Batik Image Classification : Mulwin-LBP Algorithm and Deep Neural Network," 2021. doi: 10.1109/ISMEE54273.2021.9774230.

[24] A. H. Rangkuti, A. Harjoko, and A. Putra, "A Novel Reliable Approach for Image Batik Classification That Invariant with Scale and Rotation Using MU2ECS-LBP Algorithm," Procedia Comput. Sci., vol. 179, no. 2019, pp. 863–870, 2021, doi: 10.1016/j.procs.2021.01.075.

[25] A. H. Rangkuti, J. M. Kerta, and A. H. Aslamiah, "UTILIZATION of MULWIN-LBP ALGORITHM to SUPPORT BATIK IMAGE CLASSIFICATION," J. Theor. Appl. Inf. Technol., vol. 99, no. 24, pp. 6309–6318, 2021.

[26] Farida, R. E. Caraka, T. W. Cenggoro, and B. Pardamean, "Batik Parang Rusak Detection Using Geometric Invariant Moment," 1st 2018 Indones. Assoc. Pattern Recognit. Int. Conf. Ina. 2018 - Proc., no. September, pp. 71–74, 2018, doi: 10.1109/INAPR.2018.8627000.

[27] A. E. Minarno, Y. Azhar, F. D. Setiawan Sumadi, and Y. Munarko, "A Robust Batik Image Classification using Multi Texton Co-Occurrence Descriptor and Support Vector Machine," 2020 3rd Int. Conf. Intell. Auton. Syst. ICoIAS 2020, pp. 51–55, 2020, doi: 10.1109/ICoIAS49312.2020.9081833.

[28] D. M. S. Arsa, "VGG16 in Batik Classification based on Random Forest," 2019. doi: 10.1109/ICIMTech.2019.8843844.

[29] A. Kasim, "The selection feature for batik motif classification with information gain value," 2017. doi: 10.1007/978-981-10-7242-0_9.

[30] L. Elvitaria, E. Fadzrin, A. Shaubari, and N. A. Samsudin, "A Proposed Batik Automatic Classification System Based on Ensemble Deep Learning and GLCM Feature Extraction Method," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 553–561, 2024.

[31] C. U. Khasanah, "Implementation of Data Augmentation Using Convolutional Neural Network for Batik Classification," 2020. doi: 10.1109/CITSM50537.2020.9268890.

[32] S. S. F. Ardyani, "A Web-Based for Demak Batik Classification Using VGG16 Convolutional Neural Network," Adv. Sustain. Sci. Eng. Technol., vol. 6, no. 4, p. 240406, 2024, doi: 10.26877/asset.v6i4.771.

[33] D. Made, S. Arsa, A. Agung, and N. Hary, "VGG16 in Batik Classification based on Random Forest," 2019 Int. Conf. Inf. Manag. Technol., vol. 1, no. August, pp. 295–299, 2019.

YOLO-CBAM: A Lightweight Attention-Guided Deep Learning Framework for Real-Time Road Damage Detection

Olzhas Olzhayev[1], Bakhytzhan Kulambayev[2]*, Azizah Suliman[3],
Assel Rustem[4], Almira Madiyarova[5], Batyrkhan Omarov[6]
International Information Technology University, Almaty, Kazakhstan[1, 4, 6]
Turan University, Almaty, Kazakhstan[2]
Asia Metropolitan University, Subang Jaya Campus, Malaysia[3]
Caspian University of Technology and Engineering named after Sh.Yessenov, Aktau, Kazakhstan[5]
Khalel Dosmukhamedov Atyrau University, Atyrau, Kazakhstan[6]
Narxoz University, Almaty, Kazakhstan[6]

***Abstract*—Accurate and real-time assessment of road infrastructure is critical for smart city maintenance and transportation safety. However, conventional object detection models often struggle with complex environmental factors, such as varying illumination, shadows, and background noise, leading to false detections and missed fine-grained defects. In this study, we propose YOLO-CBAM, a lightweight and fast neural network architecture tailored for real-time road surface damage detection. The standard YOLO11s backbone is enhanced through the integration of a Convolutional Block Attention Module (CBAM), which synergistically applies channel and spatial attention mechanisms. This modification enables the network to actively suppress irrelevant background visual noise and focus exclusively on structural defects like longitudinal cracks and potholes. Extensive experiments conducted on a comprehensive dataset reveal that the implementation of partial transfer learning significantly mitigates early-stage gradient shock, allowing the model to achieve a mean Average Precision (mAP@50) of 0.60 in just 40 training epochs. Deployed on an NVIDIA RTX 4070 Ti, the proposed framework achieves an inference speed of 25 frames per second (FPS), demonstrating an optimal balance between detection accuracy and computational efficiency. The YOLO-CBAM model provides a robust, cost-effective solution for automated video surveillance and road condition monitoring in smart city infrastructures.**

Keywords*—Computer vision; deep learning; object detection; YOLO architecture; CBAM; attention mechanism; road monitoring; transfer learning

I. Introduction

Road networks are a critical infrastructure component, ensuring economic stability and public safety. However, over time, heavy traffic and adverse weather conditions inevitably lead to the formation of defects such as longitudinal and transverse cracks, potholes, and pavement deterioration. Traditional methods of visual or manual road monitoring are time-consuming and costly and are highly susceptible to human error [1]. Therefore, city authorities are increasingly implementing automated road condition monitoring systems based on video cameras and artificial intelligence technologies [2].

In recent years, convolutional neural networks (CNNs) and deep learning methods have become the international standard for road pavement degradation analysis [3]. In particular, single-stage object detectors of the YOLO (You Only Look Once) family have become widely used due to their high image processing speed [4]. YOLO's lightweight architecture and low computational requirements make these models ideal for processing real-time video streams on mobile edge devices mounted directly in inspection vehicles [5].

Despite these advantages, accurate defect detection in real-world street conditions remains challenging [6]. Standard neural networks demonstrate high accuracy under ideal lighting conditions, but their performance drops sharply in the presence of visual noise. Tree shadows, overcast skies, glare from water, and the complex texture of the asphalt itself often reduce the predictive power of algorithms [7]. Because of this, basic YOLO models either miss fine details (e.g., fine microcracks) or produce false positives, mistaking background noise for a pavement defect.

To address the problem of noise and improve the quality of feature extraction, attention mechanisms are being actively incorporated into modern computer vision architectures [8]. One of the most effective solutions for localization problems is the Convolutional Block Attention Module (CBAM) [9]. This module acts as an intelligent spatial and channel filter: it helps the neural network adaptively recalibrate feature maps, forcing the algorithm to ignore background noise and focus exclusively on the structural boundaries of real defects [10].

The integration of attention mechanisms into the YOLO family has become an important research trend in 2024–2025. Recent studies confirm that the combination of improved YOLO backbones and attention modules (such as CBAM or SimAM) can significantly improve localization accuracy without significantly increasing the computational load [11], [12]. Furthermore, architectural optimization and the use of modern loss functions allow such hybrid networks to achieve an ideal balance between inference speed (FPS) and average accuracy (mAP) [13], [14].

*Corresponding author.

In this study, we propose YOLO-CBAM—a modified and lightweight neural network model designed specifically for real-time road surface defect detection. By integrating the CBAM module into the standard YOLO architecture, we enhance the network's ability to detect complex defects against urban textures. Furthermore, the study examines the impact of various transfer learning strategies on the stability of model optimization (including overcoming the "gradient shock" phenomenon) using the extended RDD2022 dataset [15]. The primary goal of this study is to create a fast, computationally efficient, and robust tool for automated infrastructure monitoring in smart city applications.

The remainder of this study is organized as follows. Section II presents a review of recent studies related to road damage detection, lightweight object detection frameworks, and attention mechanisms in computer vision. Section III describes the dataset preparation process, preprocessing pipeline, proposed YOLO-CBAM architecture, mathematical formulation of the attention mechanism, and optimization strategy. Section IV provides the experimental results, including quantitative evaluation, ablation studies, qualitative visualization analyses, and computational performance assessment. Section V discusses the effectiveness, practical deployment feasibility, and limitations of the proposed framework. Finally, Section VI concludes the study and outlines future research directions for improving real-time road infrastructure monitoring systems.

II. Related Works

In recent years, automated visual inspection of road surfaces has been addressed primarily using convolutional neural networks (CNNs). Single-Stage Detectors (SNDs) of the YOLO family have become particularly popular, as they provide an optimal balance between localization accuracy and inference speed [16]. Researchers are actively adapting various YOLO versions (from YOLOv5 to YOLOv8) to classify longitudinal cracks, potholes, and asphalt wear [17], [18]. While standard architectures are successful in localizing large defects, their application in real-world urban environments is often limited due to their high sensitivity to illumination changes, complex background textures, and the presence of foreign objects [19], [20]. To improve the robustness of systems, some authors propose multi-scale networks; however, their direct use without modification often leads to increased computational load, which complicates deployment in smart city video systems [21].

To address the shortcomings of standard convolutional networks, spatial and channel attention mechanisms (Attention Mechanisms) are being actively implemented in deep learning architectures [22]. These modules allow the neural network to dynamically re-evaluate the importance of different image regions, focusing on structural anomalies and suppressing background visual noise [23]. One of the most effective solutions in this area is the Convolutional Block Attention Module (CBAM) [24]. Recent studies confirm that integrating CBAM into the backbone or neck of YOLO-type detectors significantly improves the model's sensitivity to small objects, such as microcracks [25], [26]. Unlike heavier transformers, CBAM modules add virtually no additional network parameters. Experiments show that the combination of YOLO and CBAM provides a significant increase in mean average accuracy (mAP) in localizing infrastructure defects without sacrificing frame rate (FPS) [27], [28].

A key requirement for mobile road monitoring systems is the use of lightweight models capable of running on edge devices in real time [29]. For such models, weight optimization is critical. Training a network from scratch on specific datasets often results in gradient instability in early epochs [30]. Transfer learning [31] is widely used in modern literature to address this issue. Partially transferring weights pre-trained on large open datasets (e.g., COCO) allows models to avoid early gradient shock and significantly accelerates convergence [32], [33]. As noted in a number of works on infrastructure monitoring, the use of transfer learning strategies in lightweight modified networks allows achieving target accuracy indicators (mAP@50 > 0.55–0.60) in a minimum number of epochs [34], which makes such solutions highly effective for practical application.

III. Materials and Methods

A. Dataset Preparation and Preprocessing

The dataset utilized in this study is the publicly available "Road Damage" dataset sourced from the Kaggle platform. It comprises thousands of images depicting real-world asphalt degradation, categorized into critical classes such as longitudinal cracks, transverse cracks, alligator cracks, and potholes. To facilitate a comprehensive evaluation of the model's generalization capabilities and to prevent overfitting, the normalized dataset was subsequently partitioned using a randomized split. Specifically, 80% of the data was allocated to the training subset, while the remaining 20% was strictly reserved for the validation subset to benchmark the model's performance on unseen structural defects.

To ensure optimal convergence during the training phase and to comply with the architectural prerequisites of the YOLO network, a rigorous preprocessing pipeline was implemented. Initially, all diverse input images were standardized through resizing to a uniform spatial resolution of W×H=640×640 pixels. Following the spatial adjustment, pixel intensity normalization was applied to stabilize the gradient descent process and mitigate the vanishing gradient problem. The normalization converts the original 8-bit integer pixel values (ranging from 0 to 255) into a floating-point format within the range of [0,1]. This operation is mathematically defined for each pixel at coordinates (x,y) as:

$$I_{norm}(x,y) = \frac{I(x,y)}{255.0}, \quad (1)$$

Furthermore, the ground truth bounding boxes were normalized relative to the image dimensions to standardize the regression targets. For a bounding box with absolute coordinates (center x_{abs}, y_{abs}, w_{abs} and h_{abs}) the normalized coordinates are computed as:

$$x_c = \frac{x_{abs}}{W}, y_c = \frac{y_{abs}}{H}, w_{norm} = \frac{w_{abs}}{W}, h_{norm} = \frac{h_{abs}}{H}, \quad (2)$$

B. YOLO-CBAM Network Architecture

The core structural framework of the proposed detection system is derived from the YOLO11s architecture. This specific iteration of the "You Only Look Once" family was deliberately selected because it functions as a highly efficient, lightweight, single-stage object detector. It inherently offers an optimal equilibrium between high mean average precision (mAP) and the low computational overhead required for real-time edge computing deployment.

However, standard convolutional neural networks often exhibit critical vulnerabilities when deployed in uncontrolled, real-world environments. They are highly susceptible to visual distractors such as drastic changes in illumination, long evening shadows, reflective water pooling, and complex road markings. These environmental factors frequently lead to false positive detections or the omission of fine, irregular micro-cracks. To systematically eradicate this limitation, we engineered a custom modification by structurally integrating a Convolutional Block Attention Module (CBAM) into the deep feature extraction layers of the baseline YOLO architecture, thereby establishing the YOLO-CBAM model.

The holistic data flow within the proposed YOLO-CBAM architecture is visually delineated in Fig. 1. The computational process begins when the normalized image tensor is ingested by the Backbone zone. This zone is responsible for hierarchical feature extraction, aggregating low-level visual cues (such as edges and gradients) into complex, high-level semantic representations. Subsequently, these intermediate feature maps are routed through the integrated CBAM layers. The attention module acts as a dynamic, intelligent filter that recalculates feature weights, forcing the network to heavily prioritize the structural morphology of road defects while actively dampening the activation of background visual noise. Finally, the optimized feature maps are propagated through the Neck region to the Detection Head, which computes the final regression of bounding box coordinates and the corresponding class probability distributions.

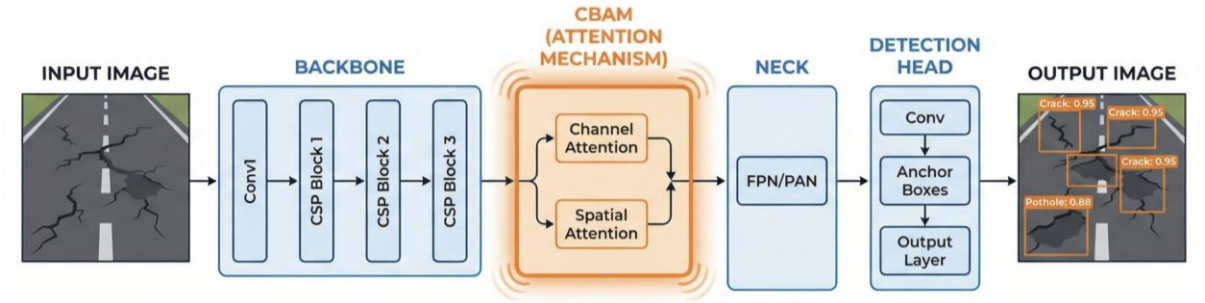


Fig. 1. Architecture of the proposed YOLO-CBAM model.

C. Mathematical Formulation of the Attention Mechanism

The Convolutional Block Attention Module is not a simple filter, but rather a sophisticated dual-attention mechanism. It sequentially refines intermediate feature maps across two distinct, complementary dimensions: the channel dimension and the spatial dimension. CBAM operates as a dual-attention mechanism that sequentially infers attention maps along channel and spatial dimensions.

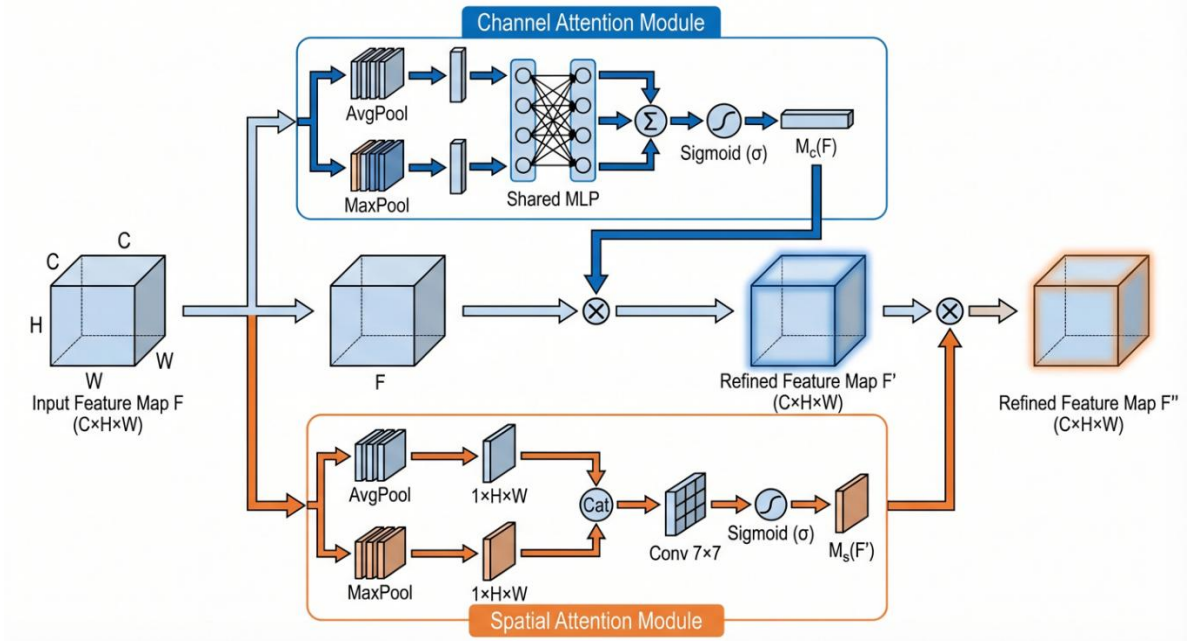


Fig. 2. Detailed workflow of the Convolutional Block Attention Module (CBAM).

As illustrated in the internal schematic in Fig. 2, let us assume the module receives an intermediate feature map denoted as $F \in R^{C\times H\times W}$, where C represents the number of channels, and H and W represent the spatial height and width. The holistic attention refinement process is mathematically formalized through sequential element-wise multiplications ($\otimes$):

$$F' = M_c(F) \otimes \mathrm{F}, \quad (3)$$

$$F'' = M_s(F') \otimes F', \quad (4)$$

here, $M_c(F) \in R^{C\times 1\times 1}$, is the 1D channel attention map $M_s(F') \in R^{C\times H\times W}$ is the 2D spatial attention map, and F'' represents the final, refined output tensor.

Channel Attention Sub-module: The primary objective of this sub-module is to determine "what" features within the map are most informative. It compresses the spatial dimensions of the input tensor F by simultaneously applying Global Average Pooling (F) and Global Max Pooling (F). These operations generate two distinct 1D spatial context descriptors. Both descriptors are then forwarded through a shared Multi-Layer Perceptron (MLP) network with one hidden layer. The output vectors are merged via element-wise summation and normalized using a sigmoid activation function (σ) to yield the channel attention weights.

$$M_c(F) = \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big), \quad (5)$$

Spatial Attention Sub-module: While the channel attention focuses on feature types, the spatial attention sub-module identifies "where" the critical structural damages are physically located on the grid. Utilizing the channel-refined feature map F', the module applies pooling operations along the channel axis. The resulting 2D maps are concatenated and subjected to a standard convolution operation ($f^{7\times 7}$) utilizing a robust 7×7 kernel. This process suppresses irrelevant background textures and highlights the morphological boundaries of the defects:

$$M_s(F') = \sigma(f^{7\times 7}([AvgPool(F'); MaxPool(F')]), \quad (6)$$

D. Loss Functions, Optimization Strategy, and Experimental Setup

To guarantee high precision in both defect classification and the exact spatial localization of bounding boxes, the network's learning process was governed by a multi-component loss function. The total loss (L) is calculated as a weighted linear combination of three distinct penalty metrics.

$$\mathrm{L_{total}} = \lambda_{\mathrm{box}}\mathrm{L_{box}} + \lambda_{\mathrm{cls}}\mathrm{L_{cls}} + \lambda_{\mathrm{dfl}}\mathrm{L_{dfl}}, \quad (7)$$

In this formulation, L_{cls} represents the classification loss, calculated using standard Cross-Entropy to penalize incorrect defect categorization. L_{box} bo denotes the bounding box regression loss (typically Complete Intersection over Union, CIoU), which mathematically evaluates the geometric overlap, central point distance, and aspect ratio alignment between the predicted boxes and the ground truth annotations. Finally, L_{dfl} stands for Distribution Focal Loss, which optimizes the fine-grained, pixel-level localization of the bounding box edges.

The empirical experiments were executed on a high-performance computational workstation operating under a Windows environment, heavily accelerated by an NVIDIA RTX 4070 Ti Graphics Processing Unit equipped with 12 GB of VRAM. The software stack included the Ultralytics and PyTorch deep learning frameworks. We implemented the AdamW optimizer with an initial learning rate set to 0.005. To effectively combat the risk of over-parameterization and model overfitting, a weight decay coefficient of 0.0005 was applied alongside a 3.0-epoch warmup period. Training was conducted over a span of 40 epochs utilizing a batch size of 32. Furthermore, a hybrid transfer learning paradigm was deployed: the standard YOLO layers were initialized with pre-trained weights to exploit generalized prior knowledge, whereas the newly integrated CBAM layers were initialized with random weights, allowing them to autonomously learn highly specific attention topographies tailored exclusively for asphalt degradation.

E. Experimental Configuration and Training Strategy

To improve model generalization capability and robustness against environmental variability, several online data augmentation strategies were applied during the training phase. The augmentation pipeline included random horizontal flipping, random scaling, HSV color-space perturbation, mosaic augmentation, and random translation operations. These transformations were dynamically applied during mini-batch generation to increase structural diversity and reduce the risk of overfitting under limited environmental conditions. Such augmentation operations are particularly important for road damage detection tasks due to frequent illumination variability, shadow interference, and heterogeneous pavement textures observed in real-world urban environments.

The optimization process utilized the AdamW optimizer in combination with a cosine annealing learning rate schedule. The initial learning rate was initialized at 0.005 and gradually reduced throughout the training epochs to stabilize convergence and improve late-stage optimization behavior. A warmup phase of 3 epochs was additionally implemented to prevent abrupt gradient instability during the early stages of training. The batch size was fixed at 32, and all experiments were conducted over 40 epochs using mixed-precision training within the PyTorch framework.

The experimental validation strategy was based on a randomized dataset partitioning protocol, where 80% of the samples were allocated for training, and the remaining 20% were reserved for validation. The validation subset remained completely isolated during training and was used exclusively for performance monitoring and model evaluation. The best-performing checkpoint was selected according to the highest validation mAP@50 metric observed during training.

It should additionally be noted that the utilized YOLO11s implementation follows the anchor-free detection paradigm introduced in recent YOLO architectures. Consequently, no manually predefined anchor box configurations were required during training. Instead, the network dynamically predicts object localization through anchor-free regression mechanisms integrated within the detection head.

IV. RESULTS

To comprehensively evaluate the efficacy of the proposed YOLO-CBAM architecture, a multifaceted analysis was conducted. This section details the quantitative metrics, qualitative visual assessments, and real-time computational performance of the model, directly comparing the custom attention-integrated network against the baseline architecture.

The analysis of the learning curves provides critical insights into how the neural network adapts its internal weights to newly introduced structural layers. Throughout the training phase, we closely monitored the behavior of the loss functions (box loss, classification loss, and distribution focal loss) alongside the mean Average Precision (mAP) metrics over the predefined 40 epochs.

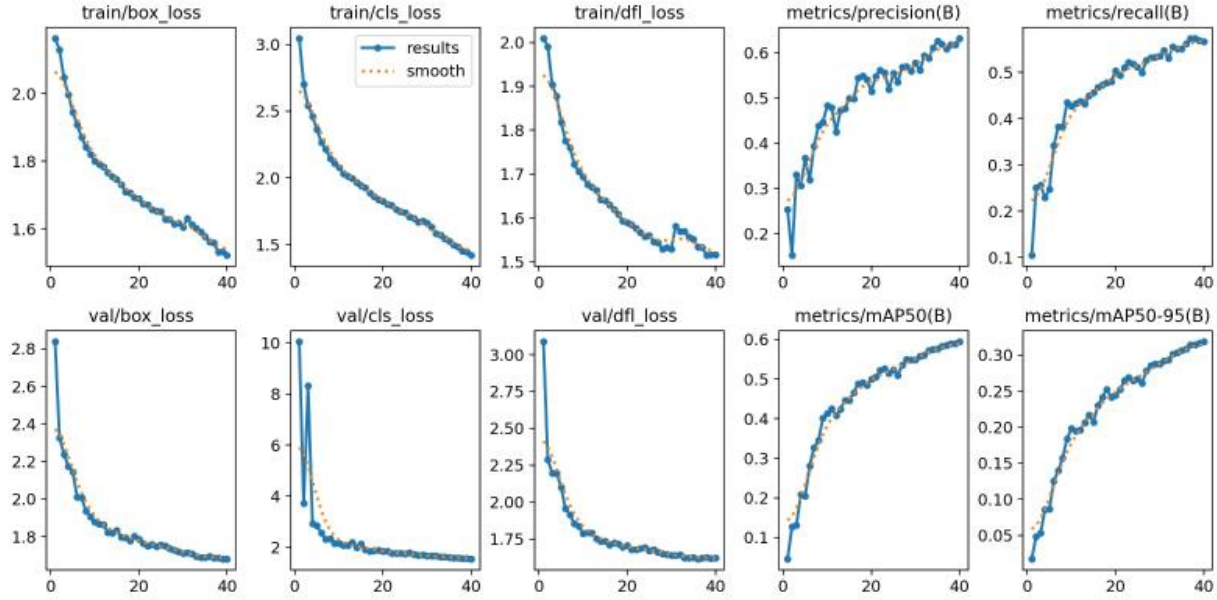


Fig. 3. Learning curves showing loss metrics and mAP over 40 epochs.

As explicitly illustrated in Fig. 3, the integration of the attention mechanism initially disrupted the learning equilibrium. During the partial transfer learning experiment, we documented a significant initial spike in the classification loss metric, where the validation class loss surged above the 10.0 threshold (val/cls_loss>10.0) within the very first epoch. In deep learning literature, this phenomenon is widely recognized as "gradient shock". This shock is a direct consequence of the architectural mismatch during initialization: the pre-trained convolutional layers of the YOLO backbone attempt to process features using established weights, while the newly appended CBAM layer, initialized with random weights, temporarily disrupts the forward pass synchronization, leading to erratic initial gradients.

Despite this severe initial perturbation, the proposed YOLO-CBAM architecture demonstrated remarkable gradient stability and structural resilience. The network rapidly overcame the gradient shock, aggressively minimizing the loss functions and reaching a stabilized error rate by the 5th training epoch. This rapid recovery period empirically confirms that the integration of the Convolutional Block Attention Module does not corrupt the underlying geometric feature extraction capabilities of the baseline YOLO architecture, but rather complements it seamlessly once the initial weights are aligned.

Beyond overall convergence, it is imperative to understand how the model differentiates between highly specific types of pavement degradation. To evaluate the exactness of the classification head, we utilized the F1-score and a normalized confusion matrix. The F1-score acts as a unified harmonic mean that perfectly balances precision and recall. The proposed YOLO11s-CBAM model, utilizing the partial transfer learning strategy, achieved a robust F1-score of 0.60. This specific value indicates a highly stable equilibrium; the network successfully minimizes false positive detections (such as misinterpreting a dark shadow as a deep crack) while simultaneously suppressing false negative occurrences (missing an actual, hazardous pothole on the road).

A deeper, class-by-class diagnostic is provided by the normalized confusion matrix presented in Fig. 4. The matrix diagonal visually confirms that the model exhibits a high degree of predictive confidence when identifying distinct, severe structural damages, particularly in the case of potholes. However, the matrix also reveals minor, localized classification overlaps. For instance, the network occasionally demonstrates slight confusion when distinguishing between visually adjacent categories, such as longitudinal and transverse cracks, which frequently share similar edge gradients depending on the camera angle. Most notably, the integration of the CBAM module profoundly impacted the background class. Compared to standard models, YOLO-CBAM drastically reduced the volume of false background triggers, demonstrating minimal confusion between actual asphalt damage and the irrelevant background surface.

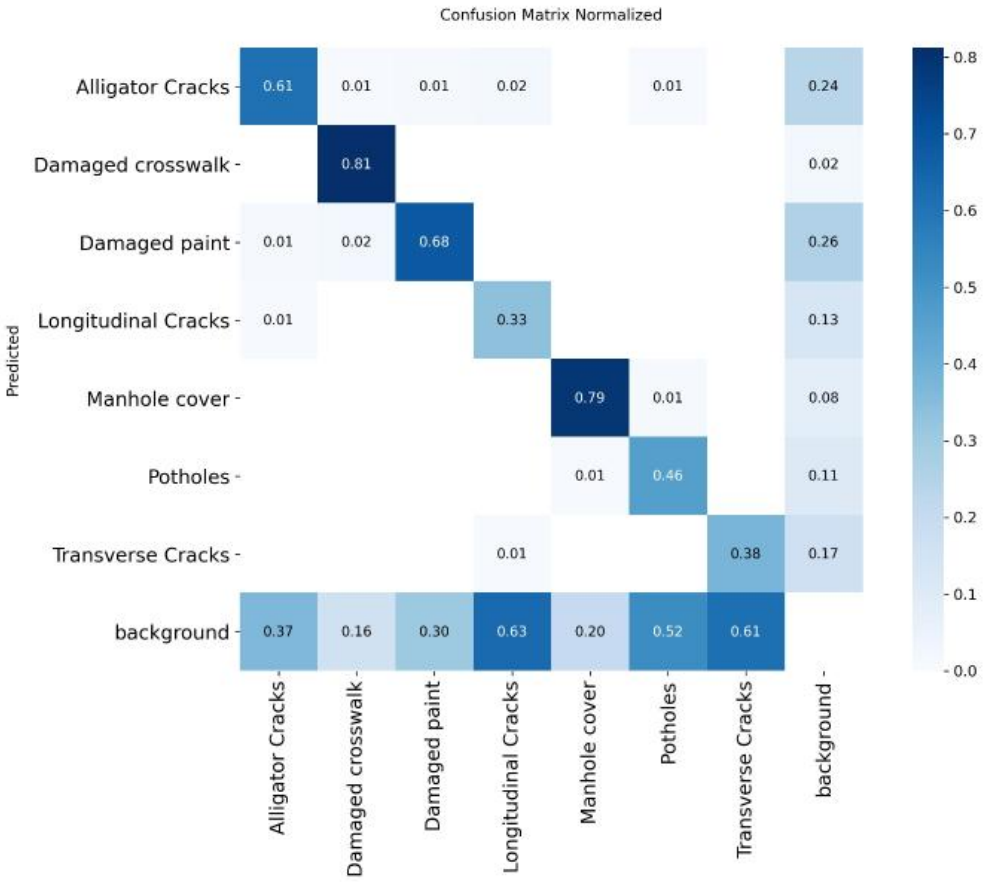


Fig. 4. Normalized confusion matrix of the YOLO-CBAM model.

To scientifically isolate and quantify the exact performance gains generated by our architectural modifications, an ablation study was performed. We conducted a direct, side-by-side comparison between the unmodified YOLO11s baseline and our enhanced YOLO-CBAM architecture. All comparative models were subjected to an identical 40-epoch training cycle using the Kaggle road damage dataset to ensure strict experimental fairness. Furthermore, we tested two separate initialization strategies for the attention module: training the entire network from scratch versus employing partial transfer learning.

The quantitative results of this rigorous ablation study are documented in Table I. The baseline YOLO11s model established a reliable foundation, achieving an mAP@50 score of 0.60. The most revealing insight emerged when the modified YOLO11s-CBAM was trained entirely from scratch, utilizing purely random weight initialization. Remarkably, even without the advantage of pre-trained knowledge, this configuration managed to achieve an mAP@50 of 0.55 in merely 40 epochs. This highly accelerated convergence rate serves as strong evidence that the CBAM module inherently possesses a superior capacity to extract critical structural features from raw asphalt imagery. Finally, when the partial transfer learning strategy was applied to the YOLO11s-CBAM model, it fully matched the baseline's overall accuracy (mAP@50 of 0.60), while simultaneously delivering marginal but important improvements in the stricter mAP@50-95 metric (0.32) and overall Recall (0.57).

While quantitative metrics provide a statistical overview, a qualitative visual assessment is required to confirm that the network is "looking" at the correct spatial coordinates. One of the most persistent challenges in automated infrastructure monitoring is environmental noise. Standard object detectors are frequently deceived by complex real-world conditions, such as footage captured during early evening hours with elongated shadows or rainy conditions that produce confusing water reflections on the asphalt.

TABLE I. IMPACT OF ARCHITECTURAL CHANGES AND INITIALIZATION STRATEGIES ON DETECTION ACCURACY

Architecture	Learning Strategy	mAP@50	mAP@ 50-95	Precision (P)	Recall (R)
YOLO11s (baseline)	Transfer Learning	0.6	0.3	0.63	0.55
YOLO11s-CBAM	Learning From Scratch	0.55	0.29	0.59	0.53
YOLO11s-CBAM	Partial Transfer Learning	0.6	0.32	0.63	0.57

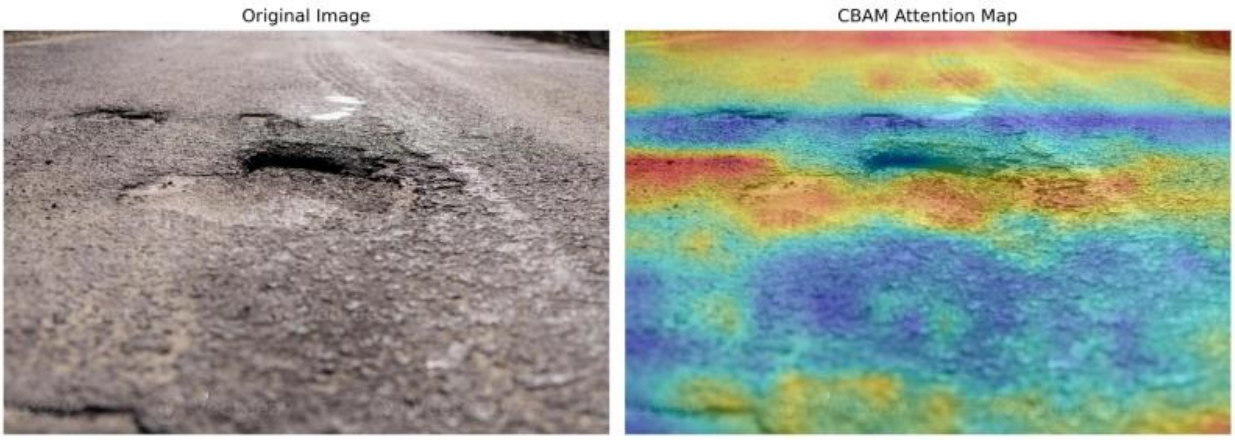


Fig. 5. Visualization of CBAM attention maps for damage detection.

To empirically verify our hypothesis, we extracted and visualized the internal attention maps generated by the network. Fig. 5 provides a direct, visual comparison between a raw, original input image containing a defect and the corresponding heat map generated exclusively by our custom CBAM layer. The heatmap visualization unequivocally demonstrates the module's effectiveness: the regions of highest neural focus, represented by intense red and yellow areas, align perfectly with the harsh structural boundaries and jagged edges of the road defect. Conversely, the surrounding undamaged road surface, including highly distracting shadows and variations in pavement color, is almost entirely ignored by the model, represented by the cool, inactive blue regions. This visual evidence conclusively proves that the dual spatial-channel attention mechanism aggressively filters out background environmental noise, forcing the convolutional layers to dedicate their computational resources exclusively to actual pavement degradation.

Achieving high accuracy is only one facet of a successful smart city solution; the model must also be practically deployable. The feasibility of deploying the YOLO11s-CBAM architecture onto mobile, edge-computing devices installed in municipal vehicles depends entirely on its inference speed and computational weight.

To simulate real-world processing capabilities, we benchmarked the model's inference speed using an NVIDIA RTX 4070 Ti GPU. Under these hardware conditions, the proposed model achieved a highly efficient average data processing speed of approximately 38.7 milliseconds per image. This translates to a continuous throughput of roughly 25 to 26 frames per second (FPS). This level of performance is more than sufficient to process high-resolution video streams in real-time, even when the cameras are mounted on vehicles traveling at high speeds through complex urban environments.

Crucially, despite the integration of the complex dual-attention CBAM layers, the overall increase in Giga Floating-point Operations Per Second (GFLOPs) was statistically negligible. The model successfully retained the "lightweight" architectural profile that is an absolute necessity for edge computing. Furthermore, the total memory footprint remained comfortably within the 12 GB VRAM limit of the testing hardware. These results indicate that the proposed architecture maintains a relatively lightweight computational profile despite the integration of the CBAM attention mechanism. Nevertheless, the current study did not include direct benchmarking on embedded edge-computing hardware platforms such as the NVIDIA Jetson series. Consequently, the suitability of the proposed framework for deployment on constrained low-power devices should be considered a prospective implementation direction rather than an experimentally validated conclusion. Future research will therefore focus on TensorRT-based optimization, quantization strategies, and empirical latency benchmarking on embedded edge platforms to evaluate real-world deployment feasibility.

V. Discussion

The experimental findings demonstrate that the integration of the Convolutional Block Attention Module into the lightweight YOLO11s architecture significantly improves the robustness of road damage detection under complex urban environmental conditions. The proposed YOLO-CBAM framework achieved stable localization performance while maintaining real-time computational efficiency, indicating that spatial-channel attention mechanisms can effectively enhance discriminative feature extraction without introducing substantial computational overhead. The obtained mAP@50 and recall metrics confirm that the attention-guided feature refinement process contributes positively to the detection of structurally irregular pavement defects, particularly in scenarios characterized by illumination variability, background clutter, and asphalt texture complexity.

Compared with conventional YOLO-based pavement monitoring systems discussed in Section II, the proposed framework demonstrates several important advantages. Earlier lightweight road inspection architectures, including YOLO-LWNet [5] and LE-YOLOv5 [17], primarily focused on reducing computational complexity through structural simplification and lightweight convolutional operations. Although these approaches achieved acceptable inference speed, their robustness under visually noisy road environments remained limited. Similarly, recent enhanced YOLOv8-based methods [6], [11], [19] improved detection sensitivity through architectural optimization; however, these models frequently

require additional computational resources due to expanded feature fusion layers and multi-scale processing modules. In contrast, the proposed YOLO-CBAM architecture introduces an efficient dual-attention refinement mechanism that selectively emphasizes defect-related structural information while suppressing irrelevant environmental activations. This property was visually confirmed through the extracted attention heat maps, where the model concentrated predominantly on defect boundaries while ignoring shadows, pavement discoloration, and reflective artifacts.

The conducted ablation experiments further validate the contribution of the CBAM integration. The experimental results indicate that even when initialized from random weights, the attention-enhanced architecture demonstrated accelerated convergence characteristics and preserved competitive localization accuracy within a relatively limited number of training epochs. Furthermore, the utilization of partial transfer learning substantially stabilized the optimization process during the early training stages by mitigating the gradient shock phenomenon associated with architectural modification. These observations are consistent with previous findings reported in transfer learning studies for infrastructure monitoring applications [31]-[34], where partially initialized lightweight detectors exhibited improved convergence stability and enhanced generalization performance.

From a deployment perspective, the proposed framework maintains the lightweight operational characteristics required for edge-computing environments and intelligent transportation systems. Despite the integration of additional attention operations, the inference speed remained within real-time processing constraints, achieving approximately 25–26 FPS on an NVIDIA RTX 4070 Ti GPU. The obtained computational efficiency suggests that the proposed framework may represent a suitable candidate for future deployment in edge-assisted smart transportation systems, although dedicated validation on embedded hardware platforms remains necessary.

Nevertheless, several limitations remain. The confusion matrix analysis revealed moderate inter-class ambiguity between visually similar crack categories, particularly longitudinal and transverse pavement degradations under oblique viewing conditions. In addition, the current dataset distribution does not comprehensively represent extreme environmental scenarios such as heavy rainfall, snow-covered pavements, severe nighttime illumination, or motion blur induced by high-speed vehicle movement. Future research will therefore focus on expanding dataset diversity, integrating advanced multi-scale contextual attention mechanisms, and investigating domain generalization strategies to improve cross-environment robustness. Additional optimization through quantization and TensorRT-based acceleration will also be explored to facilitate deployment on low-power embedded edge devices such as NVIDIA Jetson platforms.

VI. Conclusion

In summary, this study successfully developed the YOLO-CBAM architecture as a fast and automated solution for detecting road surface damage. By adding the Convolutional Block Attention Module to the YOLO11s network, the model gained the ability to clearly focus on asphalt defects while efficiently ignoring background noise like shadows and road markings. Even with an initial gradient shock, using a partial transfer learning strategy helped the network train quickly and stably. As a result, the model achieved a solid mean Average Precision (mAP@50) and F1-score of 0.60 in just 40 epochs. Importantly, the network remains lightweight, making it highly suitable for real-time edge computing on mobile devices. Moving forward, future research will expand the dataset to include extreme weather conditions and improve the detection of tiny micro-cracks. Ultimately, the YOLO-CBAM model provides a reliable, cost-effective, and powerful tool for smart city infrastructure monitoring and autonomous road maintenance.

References

[1] Safyari, Y., Mahdianpari, M., & Shiri, H. (2024). A review of vision-based pothole detection methods using computer vision and machine learning. *Sensors*, 24(17), 5652.

[2] Guerrieri, M., Parla, G., Khanmohamadi, M., & Neduzha, L. (2024). Asphalt pavement damage detection through deep learning technique and cost-effective equipment: A case study in urban roads crossed by tramway lines. *Infrastructures*, 9(2), 34.

[3] Fan, L., Wang, D., Wang, J., Li, Y., Cao, Y., Liu, Y., ... & Wang, Y. (2023). Pavement defect detection with deep learning: A comprehensive survey. *IEEE Transactions on Intelligent Vehicles*, 9(3), 4292-4311.

[4] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[5] Wu, C., Ye, M., Zhang, J., & Ma, Y. (2023). YOLO-LWNet: A lightweight road damage object detection network for mobile terminal devices. *Sensors*, 23(6), 3268.

[6] Zhang, S., Liu, Z., Wang, K., Huang, W., & Li, P. (2025). OBC-YOLOv8: an improved road damage detection model based on YOLOv8. *PeerJ Computer Science*, 11, e2593.

[7] Wang, Z., Abbas, M., & Wang, L. (2024). An attention-based improved YOLOv8 method for pavement distress detection. *In Transportation Research Board Annual Meeting*, 103.

[8] Jegadeesan, R., Beno, A., Manikandan, S. P., Rao, D. N. M., Narukullapati, B. K., Kumar, T. R., ... & Batu, A. (2022). Stable Route Selection for Adaptive Packet Transmission in 5G-Based Mobile Communications. Wireless Communications and Mobile Computing, 2022(1), 8009105.

[9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *In Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).

[10] Liang, H., Lee, S. C., & Seo, S. (2022). Automatic recognition of road damage based on lightweight attentional convolutional neural network. *Sensors*, 22(24), 9599.

[11] Li, Y., Yin, C., Lei, Y., Zhang, J., & Yan, Y. (2024). RDD-YOLO: road damage detection algorithm based on improved you only look once version 8. *Applied Sciences*, 14(8), 3360.

[12] Tao, X., Zeng, Z., Zeng, J., Zhao, T., & Dai, Q. (2024). Pavement crack detection and identification based on improved YOLOv8. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 18(1), 1-20.

[13] Pham, V., Ngoc, L. D. T., & Bui, D. L. (2024, December). Optimizing YOLO architectures for optimal road damage detection and classification: A comparative study from YOLOv7 to YOLOv10. *In 2024 IEEE International Conference on Big Data (BigData)* (pp. 8460-8468). IEEE.

[14] Wang, W., Yu, X., Jing, B., Tang, Z., Zhang, W., Wang, S., ... & Yang, L. (2025). YOLO-RD: A road damage detection method for effective pavement maintenance. *Sensors*, 25(5), 1442.

[15] Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Sekimoto, Y. (2024). RDD2022: A multi‐national image dataset for automatic road damage detection. *Geoscience Data Journal*, 11(4), 846-862.

[16] Guo, G., & Zhang, Z. (2022). Road damage detection algorithm for improved YOLOv5. *Scientific reports*, 12(1), 15523.

[17] Diao, Z., Huang, X., Liu, H., & Liu, Z. (2023). LE‐YOLOv5: A Lightweight and Efficient Road Damage Detection Algorithm Based on Improved YOLOv5. *International journal of intelligent systems*, 2023(1), 8879622.

[18] Shaghouri, A. A., Alkhatib, R., & Berjaoui, S. (2021). Real-time pothole detection using deep learning. *arXiv preprint arXiv*:2107.06356.

[19] Wang, H., Han, X., Song, X., Su, J., Li, Y., Zheng, W., & Wu, X. (2024). Research on automatic pavement crack identification Based on improved YOLOv8. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 18(6), 3773-3783.

[20] He, S., Yuan, Y., & Yin, B. (2025). LMD_YOLO: a lightweight and efficient model for pavement defects detection. *IEEE Access*.

[21] Olzhayev, O., Kulambayev, B., Sakenkyzy, N., & Belisbek, M. (2026). A Real-Time Multi-Scale Feature Pyramid YOLO Architecture for Accurate and Deployment-Efficient Road Damage Detection. *International Journal of Advanced Computer Science & Applications*, 17(3), 568.

[22] Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R. V. V., El-Ebiary, T. Y. A. B., & Rao, V. S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model. *Alexandria Engineering Journal*, 118, 482-495.

[23] Tang, Z., Wang, H., & Chamchong, R. (2026). Enhanced YOLOv8 for efficient road damage detection with spatial-channel reconstruction and multi-scale attention. *Scientific Reports*.

[24] Huan, Z., Lu, J., Wang, Y., Luo, Y., Li, Z., & Li, X. (2025, April). Enhanced CBAM-YOLOv5s for Concrete Crack Detection in Complex Environments Using Attention Mechanism and Spatial Brightness Adjustment Algorithm. *In International Conference on Civil, Architecture and Disaster Prevention and Control* (pp. 143-157). Cham: Springer Nature Switzerland.

[25] Zhong, J., Zhu, J., Huyan, J., Ma, T., & Zhang, W. (2022). Multi-scale feature fusion network for pixel-level pavement distress detection. *Automation in Construction*, 141, 104436.

[26] Niu, Q., Han, J., Sui, Z., & Xu, F. (2025). Lightweight deep neural networks: Optimization of vehicle classification using ICBAM based on depthwise separable convolutions. *PloS one*, 20(11), e0335967.

[27] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[28] Ranyal, E., Sadhu, A., & Jain, K. (2024). Enhancing pavement health assessment: An attention-based approach for accurate crack detection, measurement, and mapping. *Expert Systems with Applications*, 247, 123314.

[29] Jing, Y., Haowei, M., Ansari, A. S., Sucharitha, G., Omarov, B., Kumar, S., ... & Alyamani, K. A. (2023). Soft computing techniques for detecting cyberbullying in social multimedia data. ACM Journal of Data and Information Quality, 15(3), 1-14.

[30] Sun, H., Chen, M., Weng, J., Liu, Z., & Geng, G. (2021). Anomaly detection for in-vehicle network using CNN-LSTM with attention mechanism. *IEEE Transactions on Vehicular Technology*, 70(10), 10880-10893.

[31] Hsieh, C. C., Jia, H. W., Huang, W. H., & Hsih, M. H. (2024). Deep learning-based road pavement inspection by integrating visual information and imu. *Information*, 15(4), 239.

[32] Pan, Q., Bao, Y., & Li, H. (2023). Transfer learning-based data anomaly detection for structural health monitoring. *Structural Health Monitoring*, 22(5), 3077-3091.

[33] Du, F. J., & Jiao, S. J. (2022). Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection. *Sensors*, 22(9), 3537.

[34] Gui, R., Sun, Q., Wu, W., Zhang, D., & Li, Q. (2023). Transfer learning for cross-scene 3D pavement crack detection based on enhanced deep edge features. *Engineering Applications of Artificial Intelligence*, 123, 106452.

Deep Learning-Based Recommender System for Arabic Content with Integrated Sentiment Analysis of User Reviews

Amani Al-Ajlan[1], Nada Alshareef[2]
Department of Computer Science-College of Computer and Information Sciences,
King Saud University, Riyadh, Saudi Arabia[1]
Department of Information Technology-College of Computer and Information Sciences,
King Saud University, Riyadh, Saudi Arabia[2]

***Abstract*—Recommender systems are widely used as an information filtering technology to automatically predict and identify a set of interesting items for users based on their needs and preferences. They are widely applied in many domains, including e-commerce, social media, education, and healthcare. Recommender systems employ various filtering approaches, such as collaborative filtering, content-based filtering, and hybrid methods. Collaborative filtering is broadly categorized into memory-based and model-based approaches. Deep learning-based recommenders are a type of model-based approach that employs neural networks to capture patterns in user preferences and item features and generate accurate and personalized recommendations. In this study, we apply deep learning-based recommender systems to the Large-Scale Arabic Book Reviews Dataset (LABR) and evaluate their performance. To improve recommendation quality, we integrate sentiment analysis of user reviews using pre-trained Arabic BERT–mini and AraBERT, enabling more accurate modeling of user preferences. The results show that the combination of deep learning techniques and sentiment analysis produces more accurate recommendations, improving user satisfaction and engagement with Arabic content.**

***Keywords*—Recommender system; deep learning; sentiment analysis; LABR dataset; Model-based collaborative filtering; pre-trained model Arabic BERT; pre-trained model AraBERT**

I. Introduction

Recommender systems are effectively used as an information filtering technique to automatically predict and identify a set of interesting items for users according to their needs and preferences. Recommender systems are widely used across various applications due to the significant benefits they offer to both users and companies. They reduce the effort required to search for relevant items, thereby saving users time. Today, recommendation systems are essential parts of most applications such as Netflix, YouTube, and TripAdvisor. Recommender systems mechanisms have been shown to enhance the efficiency and consistency of decision-making. Given the vast amount of available information, these systems address the problem of information overload by filtering the user options and recommending the most suitable items, thereby improving user experience [1], [2]. There are various filtering approaches used in recommender systems, such as content-based filtering (CBF), collaborative filtering (CF), and hybrid filtering (HF). Content-based filtering uses domain knowledge to recommend items similar to those a user has previously liked. Using user profiles and item metadata, such as item descriptions, content-based recommender systems generate personalized recommendations. This approach can be implemented either explicitly, based on user-stated preferences, or implicitly, by learning from user interactions and behavior, such as clicking on links [2]. Collaborative filtering generates recommendations based on the preferences of similar users, often referred to as "neighbors" [3]. Neighbor-based algorithms are typically used in collaborative recommender systems that model relationships either between users or between items. However, collaborative recommenders suffer from the cold-start problem, which occurs when there is insufficient data for a new user or item [4]. Various collaborative filtering methods have been proposed to generate user-desired recommendations, with memory-based and model-based approaches being the most widely used [5]. Hybrid filtering is the combination of collaborative filtering and content-based filtering to obtain the best recommendation results [6]. Recommender systems are based on methods such as nearest neighbor and matrix factorization. However, deep learning has been successfully applied in various domains, including image recognition, computer vision, speech recognition, and natural language processing, and has recently been adopted in recommender system research. To improve the quality of recommendations, many businesses have adopted deep learning-based recommender systems [1]. These systems provide a significant improvement over traditional models by effectively handling high-dimensional data and addressing complex problems [1], [7]. Moreover, deep learning models are capable of capturing non-linear relationships between users and items [6].

In this study, we propose a deep learning-based collaborative filtering algorithm, which is a type of model-based collaborative filtering algorithm that utilizes neural networks. Unlike traditional methods such as matrix factorization, deep learning techniques can model non-linear relationships, sequential behaviors, and additional data sources, including product descriptions, user reviews, and images. Consequently, the proposed system can generate more personalized recommendations.

Recently, sentiment analysis of user reviews has been integrated with recommender systems to enhance the recommendation quality and address the data-sparsity problem in traditional recommender systems [8]. In such approaches, both user ratings and textual reviews are used to improve recommendations. Sentiment analysis extracts the emotional tone of user reviews to determine user preferences. By combining sentiment analysis with traditional methods such as content-based and collaborative filtering, hybrid recommender systems can be developed that provide more relevant and personalized recommendations [9]. The systems can produce more relevant recommendations by understanding user opinions, which increases user satisfaction and engagement [10]. The cold start issue, which arises when there is insufficient data for the new item or user, can also be resolved via sentiment analysis. Even without explicit ratings, sentiment analysis can identify preferences by analyzing text data. In our study, we propose a hybrid approach that combines sentiment analysis of Arabic user reviews with deep learning-based recommender systems to improve recommendation quality and solve the cold start problem. We use the Large-Scale Arabic Book Reviews Dataset (LABR) to evaluate the performance of the proposed approach. Our system takes book reviews as input and provides the corresponding sentiments for each of the books in our dataset. Then, both user ratings from LABR and sentiment scores are used to produce more personalized recommendations. This study extends our previous work [11], in which we proposed integrating sentiment analysis of user reviews with recommender systems for Arabic content. In that work, we developed two collaborative filtering-based recommender systems: the first is based on user ratings, while the second enhances recommendations by integrating sentiment analysis from user reviews. In the sentiment analysis phase, we compared three approaches: the Arabic BERT-mini model, the AraBERT model, and the Mazajak tool. In our previous work, we employed traditional collaborative filtering methods, including Singular Value Decomposition (SVD), K-Nearest Neighbors (KNN), and Non-Negative Matrix Factorization (NMF). This study builds on that work by employing deep learning–based recommender systems. The main contributions of our research are the following: 1) Proposed a hybrid approach that combines a deep learning-based collaborative recommender system with different Arabic sentiment analysis models and compares their results. 2) Applied the Arabic BERT–mini model and AraBERT sentiment analysis model to user reviews of the LABR dataset. 3) Studied the impact of using sentiment analysis models with a deep learning-based collaborative recommender algorithm on recommendation accuracy. The remainder of this study is organized as follows: Section II reviews related work, Section III presents the methodology, Section IV discusses the results, and Section V concludes the study with future work.

II. Related Work

A. Arabic Recommender Systems

We reviewed the literature and identified several studies that investigated recommender systems for Arabic content. First, Al-Ajlan and AlShareef [12] reviewed existing studies on recommender systems for Arabic content, summarizing them based on type, domain, datasets, and sentiment analysis integration. They noted that the most recommender systems for Arabic content use collaborative filtering, and some integrate sentiment analysis, which leads to higher accuracy. The most commonly used algorithm in these studies is Singular Value Decomposition (SVD). They also reported that the largest dataset used in these studies is LABR, which contains over 63,000 book reviews. Another study by Sallam et al. [13] proposed two approaches: memory-based collaborative filtering and model-based collaborative filtering using Singular Value Decomposition (SVD). They used the LABR dataset to evaluate their proposed approach. The proposed approach improved the quality of the recommendations for Arabic content and achieved low errors in terms of RMSE and MAE. Al-Ajlan and AlShareef [11] proposed integrating sentiment analysis of user reviews with traditional collaborative filtering recommender systems to improve the quality and accuracy of recommendations. They employed memory-based, including user-user and item-item models, and model-based collaborative filtering techniques, including SVD, KNN, and NMF. The LABR dataset was used to evaluate the proposed approach. The results indicate that the recommendation quality improves when recommender systems are integrated with user review sentiment analysis. The recommender system based on SVD, combined with the Arabic BERT-mini model, achieved the best performance in terms of RMSE and MAE. Additionally, Harrag et al. [14] explored the integration of sentiment analysis with recommendation systems by predicting product ratings from users' reviews using NLP and data mining techniques. In the sentiment analysis phase, they used an SVM model with the term frequency–inverse document frequency (TF–IDF) matrix. In the recommendation phase, they applied model-based collaborative filtering using SVD. They used the Opinion Corpus Arabic (OCA) dataset to evaluate the model. The results showed that the accuracy of the system is about 85% in predicting the ratings. Another study by Srifi et al. [15] evaluated the performance of some modern recommender systems when applied to Arabic content. They applied five modern recommendation systems that were initially intended for English content (ALFM, A3NCF, PARL, CARL, and CARP) and tested them on four Amazon English datasets translated into Arabic using Google Translate. Results showed that the systems performed well on Arabic content, with accuracy comparable to their performance on English data.

B. Deep Learning-Based Recommender Systems

This section reviews prior work on deep learning-based collaborative recommender systems across various domains. Li et al. [16] proposed the first system for collaborative filtering (CF) that integrates deep learning models. They introduced a deep architecture for CF that combines deep feature learning with matrix factorization. While matrix factorization-based CF methods can effectively capture the implicit relationships between users and items, they still face challenges such as the cold-start problem and data sparsity. Li et al. [16] employed neural network algorithms such as the stacked denoising autoencoder (SDAE), which enables the learning of high-level representations. However, SDAE has several disadvantages, one of which is the high computational cost of training. To address this issue, they used a variant of SDAE known as the marginalized denoising autoencoder (mDA), which avoids high computational expense by marginalizing out the random

feature corruption and provides a closed-form solution for learning model parameters. They selected three challenging tasks—movie recommendation, book recommendation, and response prediction—to evaluate the effectiveness of their proposed mDA-CF and mSDA-CF approaches. Their experiments were conducted on four datasets: MovieLens-100K, MovieLens-1M, Book-Crossing, and Advertising. The results demonstrated that the mDA-CF and mSDA-CF approaches outperformed comparable methods in all three tasks.

To facilitate book discovery on library websites with a large collection of books, Liu and Gao [17] proposed a new recommendation algorithm that integrates deep learning with matrix factorization. Their model employs a neural network to process the user-book rating matrix, aiming to address the sparsity problem commonly found in such datasets. The rating matrix is fed into a deep matrix factorization framework to learn the hidden features of users and books, which are then used to predict user ratings. Liu and Gao [17] trained and evaluated their model on the Goodbooks-10k dataset. The results show that the proposed algorithm effectively mitigates data sparsity and performs well on real-world book datasets.

Lin and Chi [18] proposed a movie recommendation system based on collaborative filtering and neural networks. In their study, a neural network was trained to predict user ratings for movies. They used the MovieLens ml-100k dataset, which contains ratings from 943 users on 1,682 movies, along with basic user and movie information. The input features included user ID, gender, age, occupation, movie ID, and movie category. The neural network architecture consisted of two hidden layers: the first with 192 neurons using ReLU as the activation function, and the second with 300 neurons using TanH. The final predicted rating was generated by the output layer. In the evaluation, the neural network-based model achieved a Mean Absolute Error (MAE) of 0.76, with a processing time of 136 seconds.

Healthcare recommender systems provide users with mobile healthcare services that are continuously accessible. However, designing patient-centered healthcare recommender systems presents several challenges, including processing the large volume of data generated by smart devices and sensors, managing dynamic networks for real-time data transmission, and the lack of effective methods for knowledge aggregation. To address these challenges, Aujla et al. [19] proposed DLRS, a deep learning-based recommender system that leverages Software-Defined Networking (SDN) within a smart healthcare environment. For evaluation, the system was tested using a dataset that included 100 patients, 100 sensors, and 20 doctors. DLRS operates through several key steps: first, a tensor-based dimensionality reduction algorithm is applied to remove irrelevant features from the collected data; second, a decision tree-based classification method groups patient queries according to different diseases; and finally, a convolutional neural network is used to generate personalized health recommendations. The system was evaluated through a case study using multiple performance metrics, including dimensionality reduction ratio, approximation ratio, accuracy, RMSE, MAPE, and latency. The results show that DLRS outperforms existing systems in its category.

III. Material and Method

A. Dataset

We used the Large-Scale Arabic Book Reviews (LABR) dataset, one of the largest sentiment analysis datasets for the Arabic language. It consists of over 63,257 book reviews, each rated on a scale of 1 to 5 stars [20]. These reviews were submitted by 16,486 users for 2,131 different books [20]. Table I summarizes the main characteristics of the LABR dataset. Each row in the LABR dataset represents a single book review submitted by a Goodreads user. Every row includes a user's rating on a scale of one to five, the Goodreads.com review id, the Goodreads.com user id, the Goodreads.com book id, and the text of the review as presented in Table II.

TABLE I. LABR Dataset Statistics

Number of reviews	63,257
Number of users	16,486
Avg. reviews per user	3.84
Median reviews per user	2
Number of books	2,131
Avg. reviews per book	29.68
Median reviews per book	6

TABLE II. Samples from LABR Dataset

Rating	Review id	User id	Book id	Review
5	279844173	6295076	2939264	كتاب اكثر من رائع ويستحق القراءة اكثر من مرة
5	322587552	8509913	2939264	قيم جداً . كل قرة اقرء احد ابوابه لانه هذا الكتاب لايتطلب قراءة سريعة بل دراســـة جيدة
4	510316957	3665588	2939264	لا يمر عليك يوم بدون تصفحه. أعتبره زاد المسلم اليومي

B. Methodology

In this section, we will present our approach that combines a deep learning-based collaborative filtering algorithm with sentiment analysis of users' reviews. Our study builds on our earlier work [11] that integrated sentiment analysis of user reviews into traditional collaborative filtering. The proposed approach consists of four main phases: data preprocessing, sentiment analysis of users' reviews, combining the sentiment score with the user rating score, and applying deep learning-based collaborative recommender systems, as presented in Fig. 1.

1) Preprocessing and cleaning: The dataset was preprocessed and cleaned using "Re" and "Pandas" Python libraries [21], [22]. The preprocessing step included normalizing Arabic characters and removing repeated letters, punctuation marks, and diacritics from the text. Then, the dataset was divided into positive, negative, and neutral categories based on the rating. Books with 4 or 5 stars were considered positive, those with 3 stars were neutral, and those with 1 or 2 stars were categorized as negative.

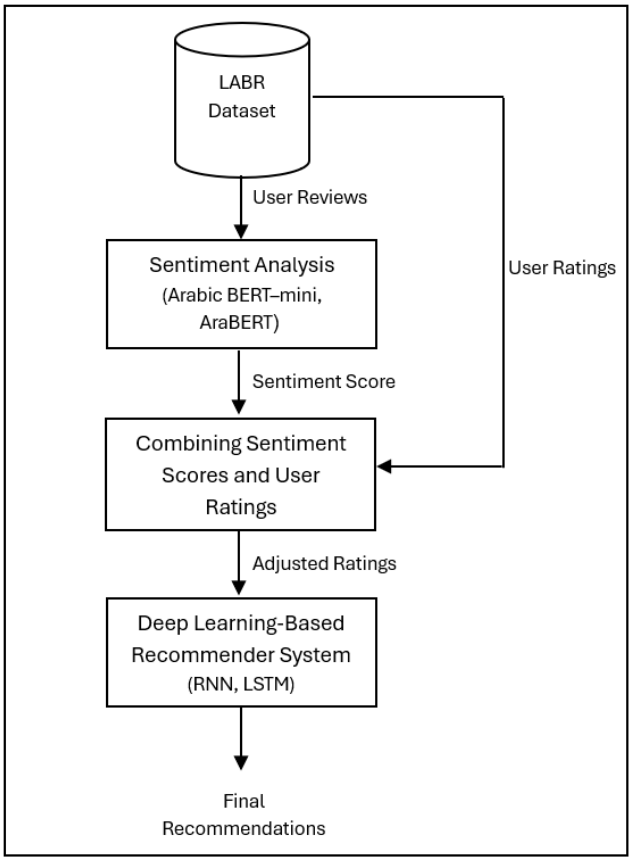


Fig. 1. The proposed approach.

2) Sentiment analysis: we applied two sentiment-analysis approaches using pre-trained Arabic language models: Arabic BERT–mini model and AraBERT [23], [24], [25]. The Arabic BERT–mini model [24], [26] was originally trained on a large Arabic corpus of about 8 billion words from OSCAR [27], Arabic Wikipedia [28], and other large text sources (around 95 GB). We fine-tuned this model using the Arabic Sentiment Twitter Corpus [29] which contains 58K Arabic labeled tweets. We also experimented with several additional datasets—SS2030 [30], the 100K Reviews dataset [31], and ArSAS [32]—and found that the Arabic Sentiment Twitter dataset produced the best performance. After tuning, the model was used to classify sentiment in the LABR dataset. Similarly, AraBERT, which is built on Google's BERT-Base architecture, was pre-trained on about 8.6 billion Arabic words from OSCAR [27], Arabic Wikipedia [28], OSIAN [33], 1.5B words Arabic Corpus [34], and Assafir news. We also fine-tuned AraBERT using the Arabic Sentiment Twitter Corpus [29] to adapt it for sentiment classification. The goal of this phase is to extract sentiment scores from user reviews in the LABR dataset.

3) Combined sentiment score with user rating score: we generate a new adjusted rating that reflects both the explicit user rating and the sentiment score extracted from the written review. Instead of relying only on the numerical rating, our method integrates it with the sentiment score produced by the sentiment analysis model. The adjusted rating is computed as follows [11]:

$$Adjusted\ Rating = \frac{(LABR\ rating + Sentiment\ rating)}{(Scale\ of\ LABR\ rating + Scale\ of\ sentiment\ rating)} \times 5 \quad (1)$$

where, the LABR rating is the original rating on a 5-point scale, and the sentiment rating is derived from the review text on a 3-point scale.

The two components are combined by normalizing their sum using the total scale and then rescaling the result to a 5-point system. This produces an adjusted rating that reflects both the explicit rating and the sentiment expressed in the review. For example, if a user assigns a rating of 5 (on a 5-point scale), and the sentiment analysis produces a score of 1 (on a 3-point scale), the adjusted rating is calculated as:

$$\frac{5+1}{5+3} \times 5 = 3.75$$

The formula assigns equal importance to both components, motivated by their complementary roles: the numerical rating captures explicit user feedback, while the sentiment score reflects implicit opinion expressed in text. Since there is no clear evidence that one component is more important than the other, both are assigned equal weight, providing a balanced baseline. Although alternative weighting schemes could be explored, this study adopts equal weighting, leaving adaptive weighting strategies for future work.

4) Deep learning-based collaborative recommender systems: the main goal of this phase is to use the adjusted ratings from the previous step, which combines sentiment analysis of user reviews and user ratings from the LABR dataset to enhance the recommendations. We developed two deep learning-based collaborative recommender systems: one using a Recurrent Neural Network (RNN) and the other using a Long Short-Term Memory (LSTM) network to learn sequential patterns in user–item interactions. In deep learning, we used "Keras" and "TensorFlow" libraries to develop deep learning model architectures. Additionally, we compared the two deep learning-based collaborative recommender systems using user ratings only and user ratings with sentiment analysis of user reviews. Also, we compare and evaluate the performance of these methods with memory-based collaborative recommender systems.

IV. Results and Discussion

A. Evaluation Metrics

Our proposed models were evaluated by using rating prediction, which measures how close the recommender's estimated ratings are to the actual user ratings. We used two metrics for evaluating the accuracy of predicted ratings: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [35]. The advantage of using RMSE over MAE is that it penalizes the error more when the difference is high (note that RMSE is always greater than or equal to MAE).

Root Mean Squared Error (RMSE) is a widely used metric because it penalizes large errors when the predicted rating is far from the actual value and less when the predictions are reasonably close. The formula for calculating RMSE is shown below.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

where, y_i = actual value, $\hat{y}_i$ = predicted value, n = number of observations.

Mean Absolute Error (MAE) is the average absolute difference between the predicted and actual ratings. The objective is to minimize the value of MAE. The formula for calculating MAE is shown below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

here, y_i = actual value, $\hat{y}_i$ = predicted value, n = number of observations

B. Experimental Results

We compared the results of the memory-based and deep learning-based collaborative filtering algorithms, which are types of model-based collaborative filtering algorithms.

1) Memory-based models: We computed both user-user and item-item similarities using two common methods, Cosine similarity and Euclidean distance. Cosine similarity measures the angle between user preference vectors, whereas Euclidean distance measures the geometric distance between them in the feature space. These techniques enable the system to generate recommendations by identifying users or items with similar behavior or preference patterns. The performance of these memory-based models is summarized in Table III.

TABLE III. MEMORY-BASED MODELS RESULTS

Models	Similarity metrics	
	Cosine similarity	*Euclidean distance*
User-User	When K= 25 RMSE = 1.86 MAE = 3.48	When K=75 RMSE = 1.90 MAE = 3.62
Item-Item	**When K= 25** **RMSE = 1.15** **MAE = 1.33**	When K=125 RMSE = 1.18 MAE = 1.39

We found that item-based models perform better and faster than user-based models when we compared the results of the two models. Moreover, Cosine similarity gave better results because it focuses on shared rating patterns between users or items and is less sensitive to rating scale differences, whereas Euclidean distance takes all dimensions into account.

2) Deep learning-based models: We developed two deep learning-based collaborative recommender systems: one using RNN and the other using LSTM. Table IV shows the performance of deep learning-based recommendation systems. As presented in Table IV, the LSTM model outperformed the RNN model as well as the memory-based collaborative filtering models.

TABLE IV. DEEP LEARNING-BASED MODELS RESULTS

Models	Performance
RNN	RMSE = 1.14 MAE = 0.92
LSTM	**RMSE = 1.07** **MAE = 0.88**

3) Memory-based models with Arabic BERT–Mini model and AraBERT: We combine sentiment analysis of user reviews with traditional memory-based recommenders. The Item-Item model with Cosine similarity metrics gave better results than the User-based model when combined with the *Arabic BERT–Mini* model as presented in Table V.

TABLE V. MEMORY-BASED WITH ARABIC BERT–MINI MODEL RESULTS

Models	Cosine similarity
User-User with Arabic BERT–mini model	When K= 25 RMSE = 1.88 MAE = 3.56
Item-Item with Arabic BERT–mini model	**When K= 25** **RMSE = 1.14** **MAE = 1.30**

Also, the Item-Item model with Cosine similarity metrics outperformed the User-based model when combined with the AraBERT model as presented in Table VI.

TABLE VI. MEMORY-BASED WITH ARABERT MODEL RESULTS

Models	Cosine similarity
User-User with AraBERT model	When K= 25 RMSE = 1.90 MAE = 3.63
Item-Item with AraBERT model	**When K= 25** **RMSE = 1.15** **MAE = 1.34**

4) Deep learning-based models with Arabic BERT–Mini model and AraBERT: We combine sentiment analysis of user reviews with deep learning-based recommender systems. The RNN achieved better results than LSTM when combined with the Arabic BERT–Mini model as presented in Table VII.

TABLE VII. DEEP LEARNING-BASED MODELS WITH ARABIC BERT–MINI MODEL RESULTS

Models	Performance
RNN with Arabic BERT–mini model	**RMSE = 0.89** **MAE = 0.83**
LSTM with Arabic BERT–mini model	RMSE = 1.00 MAE = 0.89

Additionally, the RNN model integrated with AraBERT outperforms the LSTM model with AraBERT, as presented in Table VIII.

TABLE VIII. DEEP LEARNING-BASED MODELS WITH ARABERT MODEL RESULTS

Models	Performance
RNN with AraBERT model	**RMSE = 0.89** **MAE = 0.83**
LSTM with AraBERT model	RMSE = 1.02 MAE = 0.83

C. Discussion

The results show the usefulness of integrating the sentiment analysis of user reviews into the recommender system to produce better recommendations. Our proposed approach takes into account two aspects: rating and sentiment analysis, which lead to better recommendations. In addition, the experimental results demonstrate that deep learning-based approaches outperform traditional memory-based models. In particular, the RNN and LSTM models achieved lower RMSE and MAE values compared to User-User and Item-Item methods, confirming their ability to capture complex user-item interactions. Among the deep learning models, RNN showed slightly better performance than LSTM, especially when integrated with Arabic BERT–mini and AraBERT models, where it achieved the best results (RMSE=0.89, MAE=0.83). This improvement highlights the effectiveness of incorporating contextual semantic features derived from sentiment analysis of user reviews. Furthermore, the integration of sentiment analysis significantly enhanced the recommendation accuracy, demonstrating that leveraging textual review information provides richer user preference representation compared to rating-only approaches. However, despite their superior performance, deep learning models require large datasets and substantial computational resources to achieve optimal results. This limitation may affect their applicability due to limited data availability.

Compared to traditional recommendation approaches, such as those reported in [13], that rely primarily on numerical ratings and do not incorporate sentiment information from user reviews, our proposed models achieved lower error rates, particularly when integrating sentiment-aware features. This confirms that incorporating semantic information from user reviews provides a significant advantage over rating-based methods alone.

This study extends our previous work [11] by evaluating RNN and LSTM-based models. However, it does not include a direct comparison with traditional collaborative filtering methods, such as matrix factorization or neighborhood-based approaches. In our previous work [11], we evaluated conventional filtering methods, including SVD, KNN, and NMF, and achieved better performance, as indicated by lower RSME, using the same dataset and sentiment analysis approach. This can be attributed to the fact that deep learning models typically require large amounts of data for effective training, whereas the LABR dataset contains a limited number of reviews per user. In general, both studies confirm that incorporating sentiment analysis of user reviews improves recommendation quality. However, deep learning techniques often require large datasets and significant computational resources to achieve good performance. Additionally, the lack of large-scale Arabic datasets remains a challenge, and future work should focus on creating new datasets to support the development of effective recommender systems for Arabic content.

V. CONCLUSION

We proposed a deep learning-based recommender system enhanced with sentiment analysis of Arabic user reviews. The results show that deep learning models outperform traditional memory-based approaches, with the RNN integrated with Arabic BERT–mini and AraBERT models achieving the best performance. These findings highlight the effectiveness of combining semantic information with ratings to improve accuracy and personalization in recommender systems. The main contribution of this work is the integration of sentiment-aware features into a deep learning-based recommender system for Arabic content, providing a more comprehensive representation of user preferences. Despite these improvements, the reliance on large datasets and high computational resources remains a limitation. Future work will focus on developing more efficient models and evaluating them on larger datasets. Most existing recommender systems research has focused on English or other common languages. However, Arabic presents additional challenges due to its complex grammar and rich vocabulary. Therefore, further research and better support for Arabic datasets are needed to improve recommendation quality.

REFERENCES

[1] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," ACM computing surveys (CSUR), vol. 52, no. 1, pp. 1–38, 2019.

[2] A. Al-Ajlan, S. Alabdulwahab, L. Aljeraisy, A. Althakafi, and R. Alhassoun, "Tourism Group Recommender System," International Journal of Computer Science and Network Security, vol. 2022, no. 4, pp. 637–644, 2022.

[3] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in Recommender systems handbook, Springer, 2010, pp. 1–35.

[4] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data," Expert Systems with Applications, vol. 149, p. 113248, 2020.

[5] B. Alhijawi and Y. Kilani, "A collaborative filtering recommender system using genetic algorithm," Information Processing & Management, vol. 57, no. 6, p. 102310, 2020.

[6] R. Mu, "A survey of recommender systems based on deep learning," Ieee Access, vol. 6, pp. 69009–69022, 2018.

[7] S. Lee and D. Kim, "Deep learning based recommender system using cross convolutional filters," Information Sciences, vol. 592, pp. 112–122, 2022.

[8] C. N. Dang, M. N. Moreno-García, and F. D. la Prieta, "An approach to integrating sentiment analysis into recommender systems," Sensors, vol. 21, no. 16, p. 5666, 2021.

[9] N. Darraz, I. Karabila, A. El-Ansari, N. Alami, and M. El Mallahi, "Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems," Expert Systems With Applications, vol. 261, p. 125533, 2025.

[10] S. Nabil, J. El Bouhdidi, and M. Y. Chkouri, "Enhancing Recommender Systems Using Sentiment Analysis: Addressing Cold Start Issues and Improving Recommendation Quality," vol. 16, no. 4, 2025.

[11] A. Al-Ajlan and N. Alshareef, "Recommender system for arabic content using sentiment analysis of user reviews," Electronics, vol. 12, no. 13, p. 2785, 2023.

[12] A. Al-Ajlan and N. AlShareef, "A Survey on Recommender System for Arabic Content," in 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Cairo, Egypt: IEEE, Mar. 2022, pp. 316–320. doi: 10.1109/ICCI54321.2022.9756112.

[13] R. Sallam, M. Hussein, and H. Mousa, "An Enhanced Collaborative Filtering-based Approach for Recommender Systems," International Journal of Computer Applications, vol. 176, no. 41, Jul. 2020.

[14] F. Harrag, A. S. Al-Salman, and A. Alquahtani, "Arabic Opinion Mining Using a Hybrid Recommender System Approach," arXiv:2009.07397 [cs], Sep. 2020, Accessed: Aug. 13, 2021. [Online]. Available: http://arxiv.org/abs/2009.07397

[15] M. Srifi, A. Oussous, A. Ait Lahcen, and S. Mouline, "Evaluation of recent advances in recommender systems on Arabic content," J Big Data, vol. 8, no. 1, p. 35, Feb. 2021, doi: 10.1186/s40537-021-00420-2.

[16] S. Li, J. Kawale, and Y. Fu, "Deep Collaborative Filtering via Marginalized Denoising Auto-encoder," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, in CIKM '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 811–820. doi: 10.1145/2806416.2806527.

[17] A. Liu and J. Gao, "Book recommendation algorithm based on deep learning," International Journal of Science, vol. 6, no. 10, pp. 152–156, 2019.

[18] C.-H. Lin and H. Chi, "A novel movie recommendation system based on collaborative filtering and neural networks," in International Conference on Advanced Information Networking and Applications, Springer, 2019, pp. 895–903.

[19] G. S. Aujla et al., "DLRS: deep learning-based recommender system for smart healthcare ecosystem," in ICC 2019-2019 IEEE international conference on communications (ICC), IEEE, 2019, pp. 1–6.

[20] M. Aly and A. Atiya, "LABR: A Large Scale Arabic Book Reviews Dataset," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Aug. 2013, pp. 494–498.

[21] "pandas - Python Data Analysis Library." Accessed: Aug. 12, 2025. [Online]. Available: https://pandas.pydata.org/

[22] "re — Regular expression operations — Python 3.10.4 documentation." Accessed: Aug. 12, 2025. [Online]. Available: https://docs.python.org/3/library/re.html

[23] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv preprint arXiv:2003.00104, 2020.

[24] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2054–2059.

[25] "AraBERTv2 / AraGPT2 / AraELECTRA. AUB MIND." Accessed: Aug. 12, 2025. [Online]. Available: https://github.com/aub-mind/arabert

[26] "asafaya bert mini arabic model." Accessed: Sep. 12, 2025. [Online]. Available: https://metatext.io/models/asafaya-bert-mini-arabic

[27] "OSCAR." Accessed: Sep. 12, 2025. [Online]. Available: OSCAR

[28] "Wikimedia Downloads." Accessed: Sep. 12, 2025. [Online]. Available: https://dumps.wikimedia.org/backup-index.html

[29] "Arabic Sentiment Twitter Corpus." Accessed: Sep. 12, 2025. [Online]. Available: https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus

[30] "Arabic Sentiment Analysis Dataset SS2030 Dataset." Accessed: Sep. 12, 2025. [Online]. Available: https://www.kaggle.com/snalyami3/arabic-sentiment-analysis-dataset-ss2030-dataset

[31] "Arabic 100k Reviews." Accessed: Sep. 12, 2025. [Online]. Available: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews

[32] A. Elmadany, H. Mubarak, and W. Magdy, "Arsas: An arabic speech-act and sentiment corpus of tweets," Osact, vol. 3, p. 20, 2018.

[33] I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja, "OSIAN: Open source international Arabic news corpus-preparation and integration into the CLARIN-infrastructure," in Proceedings of the fourth arabic natural language processing workshop, 2019, pp. 175–182.

[34] I. A. El-Khair, "1.5 billion words arabic corpus," arXiv preprint arXiv:1611.04033, 2016.

[35] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender systems: an introduction. Cambridge university press, 2010.

A Knowledge-Enhanced Cross-Modal Transformer Network for Sentiment Analysis in Intelligent Interaction

Chunyan Huang, Xinlu Sun*
China Zhejiang Business College, Hangzhou 310053, China

***Abstract*—With the rapid advancement of multimodal emotion recognition technology, sentiment analysis models that integrate heterogeneous information—such as facial expressions and vocal intonation—are driving human–computer interaction and affective computing toward multidimensional, objective, and highly accurate approaches. Conventional emotion recognition methods typically rely on a single-modal input and therefore struggle to capture complex semantic associations and deep emotional features, which in turn undermines the stability of recognition results. A knowledge-enhanced cross-modal Transformer network (KCTN) model was proposed for sentiment analysis, which incorporates a multimodal fusion module and a long-range affective integration module to achieve deep collaborative modeling across text, speech, and facial expression features. This framework substantially enhances the completeness and robustness of emotional semantic representations. Experimental results on the self-built EC-SFED multimodal dataset and the publicly available dataset CMU-MOSI demonstrate that KCTN surpasses several mainstream baseline models in both accuracy and macro-averaged F1 score, validating its superior performance in intelligent interaction and affective computing applications.**

Keywords*—Multimodal sentiment analysis; transformer network; emotion recognition; psychological assessment; intelligent interaction

I. Introduction

Recent research on sentiment analysis within university student populations has primarily evolved along two methodological trajectories: traditional machine learning–based approaches and deep learning–driven models. Traditional methods generally require complex feature engineering on social or textual data. Wang and Zhang [1] designed a topic–sentiment hybrid model to identify students' emotional attitudes toward different discussion themes. Iram [2] proposed an automated framework for detecting affective tendencies in student-generated posts on the Facebook platform. Zhang et al. [3] adopted a sentiment analysis approach based on the Robustly Optimized BERT Pretraining Approach-Whole Word Masking (RoBERTa-WWM) capable of capturing both local semantic features and contextual associations within student discourse.

Deep learning techniques enable the automatic extraction of high-level semantic features and complex associations from multimodal data, and the complementary information across modalities enhances the effectiveness of psychological assessment. In previous studies, Song et al. [4] developed a Multi-level Attention and Multi-task (MAM) fusion model that incorporates an attention mechanism to support multitask multimodal sentiment analysis. Sun et al. [5] designed an efficient multimodal Transformer architecture featuring a dual-level feature restoration mechanism, in which global–local interactive modeling is utilized to improve multimodal fusion performance. Huan et al. [6] proposed a unified multimodal framework composed of a translation module and a prediction module, achieving structured cross-modal integration and semantic alignment. Tsai et al. [7] introduced the Multimodal Transformer (MulT) model, which employs cross-modal attention to process unaligned multimodal data and demonstrated strong performance in comparative experiments. Hazarika et al. [8] designed the Modality-Invariant and -Specific Representations (MISA) framework, which maps each modality into two distinct subspaces to enhance fusion outcomes. Han et al. [9] developed the MultiModal InfoMax (MMIM) framework, where intermodal information is maximized and jointly optimized with the primary task to strengthen modeling performance. Huddar et al. [10] proposed an attention-based multimodal contextual fusion approach that achieved notable improvements in emotion classification. Luo and Zhu [11] introduced the Multi-Dynamic Aware Network (MultiDAN), in which multilayer multi-angle interaction perception is employed to enhance multimodal language sentiment analysis. Yang et al. [12] constructed the CCIN-SA model based on a composite cross-modal interaction network to address challenges associated with multimodal fusion and interactivity, thereby improving classification accuracy. Zhong et al. [13] proposed the KET model, which dynamically integrated commonsense information by computing the cosine similarity and emotional intensity between text and commonsense knowledge features. In addition, the effectiveness of multimodal deep learning for extracting semantic and affect-related features has also been demonstrated in recent applied studies [14, 15].

In multimodal sentiment analysis for psychological assessment, various fusion strategies and deep learning models have been proposed, resulting in notable improvements in emotion recognition accuracy and robustness. However, most evaluation frameworks have primarily emphasized outcome-oriented measures while overlooking behavioral data generated during the assessment process. Intelligent interaction systems, driven by artificial intelligence technologies, create emotional experiences and produce substantial amounts of structurally

*Corresponding author.

complex behavioral process data through visual and speech modalities. These data capture the decision-making trajectory of participants and document interaction outcomes at critical points, thereby addressing the limitations of traditional psychological assessments that rely solely on textual information. This provides valuable insights into individual psychological and behavioral patterns. In this study, a multimodal fusion module and a long-range affective fusion module were integrated. The multimodal fusion module leverages heterogeneous sources, including text, speech, and video. The long-range affective fusion module aggregates affective context from subsequent utterances of the same speaker into the representation of the current utterance, effectively capturing long-distance semantic dependencies. The optimization of deep learning models to strengthen affective computing accuracy constitutes a critical pathway for developing intelligent auxiliary systems for student mental health monitoring.

II. Model Design

A. Task Definition

In the affective computing task for conversational scenarios, a set of multimodal samples is defined as Y. Each dialogue sample contains N successive utterances denoted as $U=\{u_1,u_2,...u_n\}$, and is associated with M speakers represented as $E=\{e_1,e_2,...e_m\}$. For each utterance u_i, multimodal information is composed of three modalities—text (T), audio (A), and visual (V)—such that $u_i=\{u_i^c\}$ and $c\in\{A,V,T\}$. Based on predefined emotion categories, sentiment analysis aims to infer the emotional polarity $p\in$ {negative, neutral, positive} for each speaker by jointly modeling speaker-related characteristics, contextual dependencies, and multimodal cues. Through this process, the emotion label of a new conversational sample can be accurately predicted.

B. Model Architecture

The proposed KCTN model (Fig. 1) for psychological assessment–oriented sentiment analysis incorporates the following three primary methodological contributions:

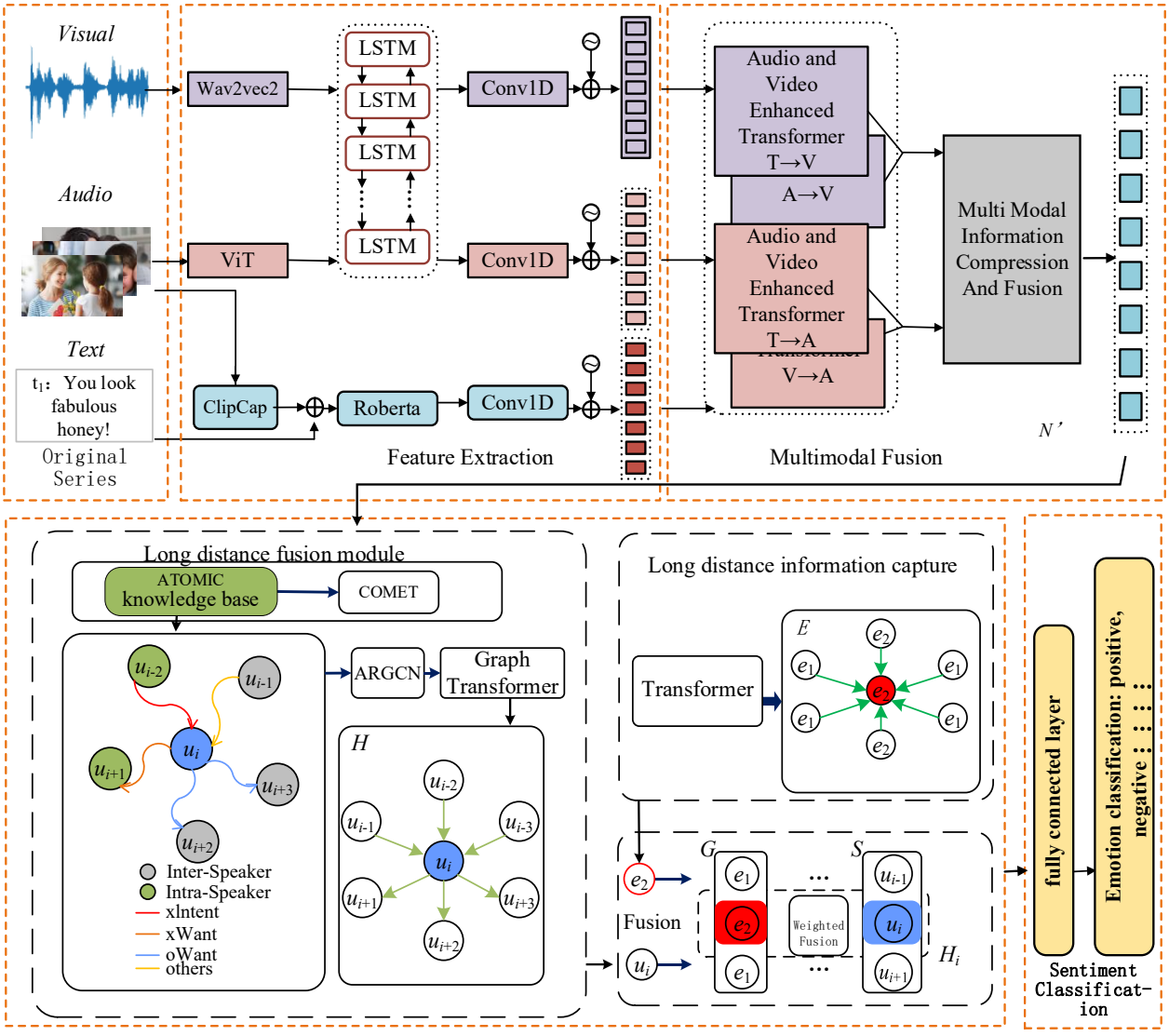


Fig. 1. Overall architecture of the proposed model.

1) In terms of feature extraction, extraction strategies for text, visual, and audio modalities were deeply studied. To address missing aspect-level emotional semantics in text, ClipCap was employed to generate natural language descriptions that align closely with image content, after which textual features were derived using the RoBERTa model. Visual features were extracted using OpenFace 2.0 and a Vision Transformer (ViT) model, while audio features were captured through Wav2Vec 2.0 and OpenSMILE.

2) In the multimodal fusion module, a cross-modal enhanced Transformer fusion network was introduced. Long Short-Term Memory (LSTM) networks, temporal convolutions, and positional embeddings were incorporated to effectively capture temporal dynamics and contextual dependencies within audio–visual modalities. Through an audio–visual enhanced Transformer module and an information compression–fusion module, non-text modalities were strengthened and deep intermodal interactions were achieved.

3) In the long-range affective fusion module, a multi-stage processing architecture was adopted. External knowledge was extracted from textual content using Commonsense Transformers (COMET) [16], and the dialogue structure was modeled through a Graph Neural Network (GNN) that transforms conversational data into graph representations to obtain sentence-level semantic features. Subsequently, the self-attention mechanism of the Transformer architecture was used to capture long-range affective dependencies across speaker utterances. Finally, a feature fusion mechanism was used to integrate sentence-level semantic representations with speaker-level affective features, thereby enhancing the model's affective computing capability.

III. Feature Extraction and Fusion

A. Text Feature Encoding

In the text feature encoding module, features from both text and images are integrated, using visual information to compensate for potentially missing affective semantics in textual data. Inspired by the multimodal processing framework proposed by Khan and Fu [17], the ClipCap image description generator was incorporated. Built upon the advanced Contrastive Language-Image Pretraining (CLIP) cross-modal pretraining framework, ClipCap integrates visual features with the prior knowledge of a language model to generate semantically precise and affectively expressive descriptions. These descriptions preserve the global emotional characteristics of the image and enable effective transfer of visual information into textual semantics.

The original text E and the image description text C were subsequently combined into a unified token sequence X through structured concatenation with special markers. A [SEP] separator was inserted between the two text segments to explicitly define modality boundaries, and a [CLS] token was placed at the beginning of the sequence to aggregate global semantic information. The resulting sequence X can be expressed below. This structured formulation preserves the fine-grained characteristics of the original text while incorporating additional semantic information generated from visual content.

$$X=[CLS]t_{E_1}t_{E_2}\ldots t_{E_{N_E}}[SEP]t_{C_1}t_{C_2}\ldots t_{C_{N_C}}[SEP] \quad (1)$$

where, Ei and Ej denote the tokens of sentences E and C, respectively. The number of tokens in E is represented as N_E, while the number of tokens in C is represented as N_C.

To fully capture the deep semantic features embedded in the textual sequence, RoBERTa-base was employed as the encoder. In the experimental configuration, the encoder was set to a 12-layer architecture with a hidden dimension of 768, and the maximum sequence length was fixed at 64 tokens. Inputs shorter than this threshold were padded with the [PAD] token, whereas overlength sequences were processed through intelligent truncation. For a given sample from the dataset, the encoded representation was obtained as follows:

$$I_t=RoBERTa\left(X;\theta^R\right)\in R^{l_t\times d_t} \quad (2)$$

where, l_t denotes the length of the text sequence, d_t denotes the dimensionality of the textual representation, and θ^R represents the parameters of the RoBERTa model.

B. Audio Feature Encoding

The multimodal emotional information generated during intelligent interaction requires the preservation of temporal continuity and contextual dependencies. However, the raw audio and visual features extracted typically consist of independent frame-wise representations that lack sufficient contextual information. To address this limitation, an LSTM network was employed to encode these modalities. In multimodal sentiment analysis for intelligent interaction, audio features are commonly categorized into three types: spectral features, timbre-related features, and prosodic features. In this study, the wav2vec2-base-960h pretrained model was utilized for feature extraction. This model was pretrained and fine-tuned on the 960-hour LibriSpeech dataset. All audio data were uniformly resampled to a 16 kHz sampling rate. New samples underwent preprocessing, feature encoding, masking, contextual representation, feature quantization, contrastive learning, loss calculation, fine-tuning, and feature extraction, resulting in the acquisition of 768-dimensional audio representations. For a given audio sequence $A=\{a_1,a_2,\ldots a_n\}$, the extracted audio feature representation is defined as:

$$I_a=Wav2vec2.0(A)\in R^{l_a\times d_a} \quad (3)$$

where, l_a denotes the audio sequence length, and d_a denotes the audio sequence dimensionality.

During the audio feature encoding process, OpenSMILE [18] with the IS10 configuration was employed to extract speech segments. Three predefined standard feature sets were included in this configuration: ComParE 2016 (a large-scale paralinguistic feature set), GeMAPS (a basic acoustic parameter set), and its extended version, eGeMAPS. The detailed configuration of these feature sets is presented in Table I.

TABLE I. OVERVIEW OF OPENSMILE FEATURE SETS

Feature Set	Number of Features	Feature Type
ComParE 2016	6373	Static features
GeMAPS	62	High-level statistical function (HSF) features
eGeMAPS	88	Extended features + HSF features

For affective computing tasks, the ComParE 2016 configuration was adopted to extract 6,373 acoustic features from each speech segment. A fully connected network was then used to reduce these features to 1,582 dimensions for the IEMOCAP dataset and to 300 dimensions for the MELD dataset; the resulting unified representation is denoted as u_t^A. To enhance temporal modeling capability, a bidirectional LSTM architecture was employed to capture contextual dependencies within the audio signal. The computation process is expressed as follows:

$$\mathbf{t}_i^{\mathrm{A}},\mathbf{H}_i^{\mathrm{A}}=\left[\overrightarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{A}},\overrightarrow{\mathbf{H}_{i-1}^{\mathrm{A}}}\right),\overleftarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{A}},\overleftarrow{\mathbf{H}_{i+1}^{\mathrm{A}}}\right)\right] \quad (4)$$

where, $\boldsymbol{H}_{i-1}^{A}$ represents the (i-1)-th hidden layer state, $\boldsymbol{H}_{i+1}^{A}$ represents the (i+1)-th hidden layer state, and $\boldsymbol{t}_i^{A}$ denotes the context-aware audio encoding.

C. Visual Feature Encoding

During the intelligent interaction process in psychological assessment, facial micro-expression dynamics were captured through video data. Visual feature extraction was performed using the OpenFace 2.0 toolkit and the ViT model. OpenFace 2.0 extracts video features through four stages: frame decomposition, facial detection and landmark localization, facial alignment, and facial feature extraction. The ViT model, built upon the Transformer architecture for image processing, divides each input frame into fixed-size patches and encodes them using a Transformer-based pipeline. The feature extraction process includes video preprocessing, selection of effective frames, image segmentation into patches, patch embedding, ViT-based encoding, and feature extraction. Through multiple layers of Transformer encoders, each token is dynamically updated to feature vectors integrating global contextual semantics with local visual details [19]. A compact representation of each video frame was then obtained by aggregating the encoded tokens. Given a preprocessed video sequence $V=\{v_1,v_2,...v_n\}$, the extracted visual feature representation is defined as:

$$I_v=ViT(V)\in\mathrm{R}^{l_v\times d_v} \quad (5)$$

where, l_v denotes the length of the video sequence, and d_v denotes the dimensionality of the video sequence.

For the video-based emotion extraction task, the raw visual features were initially obtained as independent frame-level representations. Since multimodal sentiment analysis requires the preservation of temporal continuity and contextual dependencies, a 342-dimensional visual feature, denoted as u_t^V, was first extracted. Similar to the audio encoding process, a bidirectional LSTM network was employed to capture contextual cues. The computation is defined as follows:

$$\mathbf{t}_i^{\mathrm{V}},\mathbf{H}_i^{\mathrm{V}}=\left[\overrightarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{V}},\overrightarrow{\mathbf{H}_{i-1}^{\mathrm{V}}}\right),\overleftarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{V}},\overleftarrow{\mathbf{H}_{i+1}^{\mathrm{V}}}\right)\right] \quad (6)$$

where, $\boldsymbol{H}_{i-1}^{V}$ and $\boldsymbol{H}_{i+1}^{V}$ represent the (i-1)-th and (i+1)-th hidden layer states, respectively, and $\boldsymbol{t}_i^{V}$ denotes the context-aware video encoding. To ensure that elements within each input sequence adequately capture interactions with adjacent elements, a one-dimensional temporal convolution layer was applied, which transforms the states into a unified dimension to facilitate subsequent processing. The transformation is computed as:

$$\widehat{F}_t=\mathrm{Conv1D}(I_t,\mathrm{kernel}_t)\in\mathrm{R}^{l_t\times d} \quad (7)$$

$$\widehat{F}_{\{a,v\}}=\mathrm{Conv1D}\left(I_{\{a,v\}},\mathrm{kernel}_{\{a,v\}}\right)\in\mathrm{R}^{l_{\{a,v\}}\times d} \quad (8)$$

where, $\mathrm{Kernel}_{\{t,a,v\}}$ denotes the convolution kernels corresponding to the text, audio, and visual modalities, and d represents the modality dimension.

D. Multimodal Fusion

For management units, two remote operation and maintenance modes are provided.

1) Multi-Modal Enhanced Transformer (CMET) module: In sentiment analysis, the textual modality typically plays a dominant role in conveying explicit semantic information, whereas non-textual modalities often provide auxiliary affective cues. To strengthen the emotional representational capacity of non-textual modalities, a CMET module was introduced. This module establishes a bidirectional cross-modal attention mechanism to facilitate complementary information exchange and collaborative optimization across modalities, thereby significantly enhancing the affective expressiveness of non-textual modalities. In the module, and denote the two non-textual modalities, while t denotes the textual modality. Fig. 2 shows the architecture of the audio–visual enhanced Transformer.$\alpha\beta$

The general formulation of $CMET(Q,K,V)$ is given as:

$$\mathrm{CMET}_{\mathrm{mul}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (9)$$

where, Q denotes the queries, K denotes the keys, V denotes the values, and d_k represents the dimensionality of the key or value vectors. The computation of Q, K, and V in the multi-head attention mechanism is expressed below, with the weight matrices defined as $W_{Q_\alpha}\in\mathrm{R}^{\mathrm{d}_\alpha\times d_k}$, $W_{K_\beta}\in\mathrm{R}^{\mathrm{d}_\beta\times d_k}$, and $W_{V_\beta}\in\mathrm{R}^{\mathrm{d}_\beta\times d}$:

$$Q_\alpha=Z_\alpha W_{Q_\alpha}\in\mathrm{R}^{l_\alpha\times d_k} \quad (10)$$

$$K_\beta=Z_\beta W_{K_\beta}\in\mathrm{R}^{l_\beta\times d_k} \quad (11)$$

$$V_\beta=Z_\alpha W_{V_\beta}\in\mathrm{R}^{l_\alpha\times d} \quad (12)$$

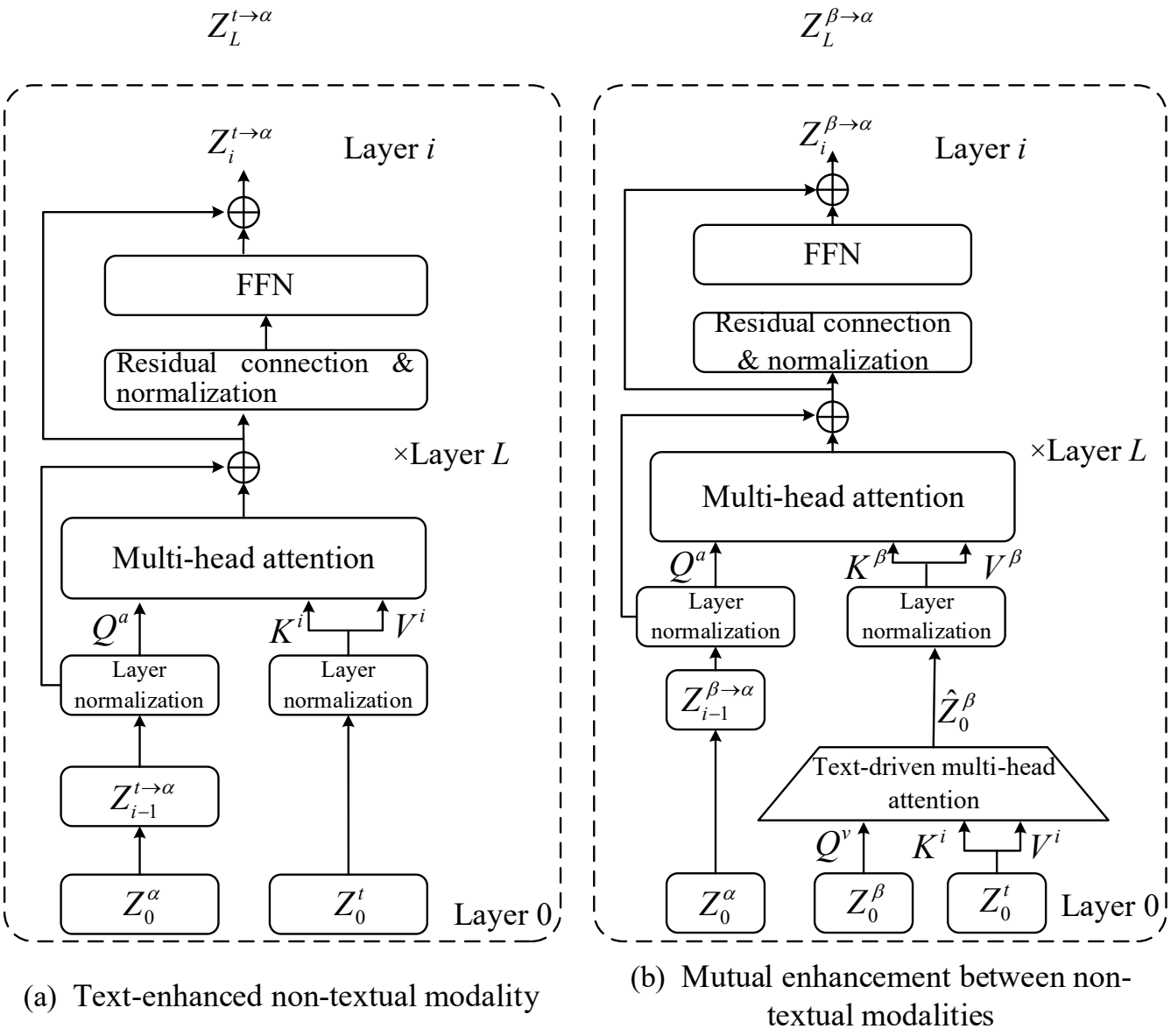


Fig. 2. Architecture of the audio–visual enhanced transformer.

For multimodal psychological testing, a cross-modal interaction encoder was designed that uses the text modality to compensate for underperforming modalities through feature enhancement: *Text→Visual*, *Audio→Visual*, *Text→Audio*, and *Visual→Audio*. This framework fosters deep inter-modal interaction between visual and audio streams while preserving information flow across all modalities. The enhancement of the visual modality is exemplified by the following formula:

$$\tilde{\mathbf{Z}}_i^{T\to V}=\mathrm{CMET}_{(i),\,\mathrm{mul}}^{T\to V}\left(\boldsymbol{t}_{(i\text{-}1)}^{T},\boldsymbol{t}_i^{T},\boldsymbol{t}_i^{V}\right) \tag{13}$$

$$\tilde{\mathbf{Z}}_i^{A\to V}=\mathrm{CMET}_{(i),\,\mathrm{mul}}^{A\to V}\left(\boldsymbol{t}_{(i\text{-}1)}^{A},\boldsymbol{t}_i^{A},\boldsymbol{t}_i^{V}\right) \tag{14}$$

The CMET module stacks multiple layers of cross-modal attention blocks, where the output of each layer serves as the input to the next, progressively enhancing audio and visual modality representations. As an example, for the enhancement of the audio modality via the visual modality (V→A), the audio representation was first initialized as $Z_0^{v\to a}=Z_0^a$, after which textual content was utilized to guide the visual modality in enriching affective information. The enhancement process is expressed as:

$$\hat{Z}_0^v=\mathrm{CMET}_{multi}^{T\to V}=\mathrm{Multihead}\left(Q_v,K_t,V_t,\right) \tag{15}$$

At layer i, the output $Z_{i-1}^{v\to a}$ from the previous layer and the text-enhanced modality information $\hat{Z}_0^v$ are integrated to compute attention weights, capturing inter-modal interactions. This integrated output is normalized to produce $\hat{Z}_i^{v\to a}$, which is then passed through a Feedforward Neural Network (FFN), yielding the final enhanced audio–visual representation $Z_i^{v\to a}$. The calculation is defined by the following formula:

$$\hat{Z}_i^{v\to a}=\mathrm{CMET}_{multi}^{v\to a}\left(\mathrm{LN}(Z_{i-1}^{v\to a}),\mathrm{LN}\left(\hat{Z}_0^v\right)\right)+\mathrm{LN}(Z_{i-1}^{v\to a}) \tag{16}$$

$$Z_i^{v\to a}=\mathrm{FFN}\left(\mathrm{LN}\left(\hat{Z}_i^{v\to a}\right)\right)+\mathrm{LN}\left(\hat{Z}_i^{v\to a}\right) \quad (17)$$

where, LN() denotes layer normalization, and FFN() denotes the feedforward neural network. In standard Transformer architectures, the ReLU activation function is commonly employed. The FFN applies an identical linear transformation and activation function to each element in the sequence independently. The FFN is formulated as:

$$\mathrm{FFN}(X)=\max(0, xW1+b1)W2+b2 \quad (18)$$

where, W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, and max(0,y) represents the ReLU activation function applied to each element to introduce nonlinearity.

2) Information compression and fusion: Token sequences were utilized as information carriers to enable efficient compression of audio–visual data. The process involves the steps below. First, the audio–visual sequence Z_m^L is concatenated with a dimension-aligned token sequence to obtain Z_{mt}^L . The combined representation incorporates information from both the original audio–visual features and their associated tokens:

$$Z_{mt}^L=\mathrm{Concat}\left(\hat{Z}_m^L,T\right) \quad (19)$$

Subsequently, a Transformer layer is applied to extract and condense the salient information contained within the audio–visual sequence. Drawing inspiration from the ViT architecture, the transformation is expressed as:

$$T_m=\mathrm{Transformer}\left(Z_m^L;\theta_m\right) \quad (20)$$

After information compression, the extracted audio and visual feature vectors are fused into a unified representation that reflects the intrinsic correlation and complementarity between the modalities. The compressed audio feature vector T_a and the compressed visual feature vector T_v are concatenated to form a joint representation T_{av}:

$$T_{av}=\mathrm{Concat}(T_a,T) \quad (21)$$

The fused representation T_{av} is then processed using a Transformer encoder to integrate comprehensive information across audio and visual modalities. The resulting representation is defined as:

$$Z_{av}=\mathrm{Transformer}\left(\hat{T}_{av}\right) \quad (22)$$

The representation Z_{av}, obtained through information compression and fusion, captures comprehensive affective cues from audio–visual data and enhances the model's ability to understand and interpret complex emotional states. After stacking *X* layers of processing, the final textual representation Z_t and the fused audio–visual representation Z_{av} are generated. Because different modality combinations contribute unequally to sentiment analysis, an adaptive weighting module was designed to quantify the contribution of each modality to the final output. The multimodal features are then aggregated through a weighted fusion mechanism to produce the final integrated output. Following a bias-free linear transformation applied to the output vectors Z_t and Z_{av}, a softmax operation is used to compute the output weights. These weights are then applied to each representation to generate the final output $\boldsymbol{N}'$, enabling more effective fusion of multimodal information and improving the accuracy of emotion prediction. The computation is defined as:

$$\boldsymbol{n}_{\mathrm{softmax}}=\mathrm{softmax}([\boldsymbol{W}_1\otimes Z_t,\boldsymbol{W}_2\otimes Z_{av}]) \quad (23)$$

$$\boldsymbol{N}'=\boldsymbol{n}_{\mathrm{softmax}}\odot Z_t+\boldsymbol{n}_{\mathrm{softmax}}\odot Z_{av} \quad (24)$$

E. Long-Range Affective Fusion Module

The long-range affective fusion module employs a hierarchical architecture to integrate emotional information. External knowledge is first extracted using the COMET model, which is pretrained on the ATOMIC knowledge graph. COMET generates nine categories of external knowledge features associated with the input text, including speaker intent (xIntent), speaker need (xNeed), speaker attribute (xAttr), effects on the speaker (xEffect), speaker desire (xWant), speaker emotional reaction (xReact), effects on others (oEffect), others' desire (oWant), and others' emotional reaction (oReact). Next, a graph-structured representation of the dialogue is constructed using a GNN to capture local, sentence-level semantic features. A Transformer encoder is then used to model dependencies and extract affective features across distant utterances. Finally, a feature fusion mechanism is designed to adaptively integrate the local sentence-level representations with global speaker-level emotional information.

Within the long-range fusion module, a directed graph is constructed to represent the interaction flow across dialogue segments. The output from the multimodal fusion module, $\boldsymbol{N}'$, is used as the initial node feature h_i. Knowledge representations generated from COMET, pre-trained on the ATOMIC knowledge graph, are aligned with relation types. The Relational Graph Convolutional Network (RGCN) updates the hidden states of nodes through the neighboring node representations based on edge types [20], while the Attention-based Relational Graph Convolutional Network (ARGCN) extends RGCN by incorporating distance-aware attention mechanisms to further enhance representational capacity [21]. The node calculation process is expressed as:

$$\boldsymbol{h}_i'=\sigma\left(\sum_{r\in R}c_{i,j}\sum_{j\in N_i^r}x_{ij,r}+\boldsymbol{W}_4\boldsymbol{h}_i\right) \quad (25)$$

where, σ denotes the activation function, R represents the set of relation types, $h_i=N'$, N_i^r is the set of neighboring nodes connected to node v_i under relation $r\in R$, W_4 is a learnable parameter matrix, and $C_{i,j}$ denotes the normalization coefficient. To propagate the interaction information aggregated by the ARGCN, a GraphTransformer model consisting of L layers is employed. The updating operation is defined as:

$$\mathbf{h}_i^{(l+1)}=\left(1-\beta_i\right)\left(\sum_{j\in N(i)}\alpha_{i,j}\,m_j\right)+\beta_i\mathbf{W}_6\mathbf{h}_i^{(l)} \quad (26)$$

where, $N(i)$ denotes the set of source nodes connected to node i, m_j represents the message from source nodes, $\alpha_{i,j}$ is the attention score, β_i is the gating parameter for the residual connection, and W_6 denotes the mapping weight. This update rule yields the final representation for all nodes in the dialogue graph. In parallel, a Transformer network is used to capture

self-dependencies across adjacent utterances from the same speaker. The resulting speaker-level contextual representation is denoted as p_i. A function $Y_i=f(L_{past},p_i)$ aggregates information from the preceding n_{past} utterances, where L_{past} denotes the information on n_{past} utterances. For the i-th utterance in the dialogue, its feature representation $\boldsymbol{E}_i$ is computed using ARGCN. In addition, U_λ denotes the information on all utterances in the dialogue.

$$E_i=\text{ARGCN}(\boldsymbol{E}_i,\boldsymbol{E}_j),j\in[i,|U_\lambda|] \tag{27}$$

Finally, a Transformer integrates Y_i with $\boldsymbol{E}_i$. The attention scores are assigned weights, generating the final fused representation H_i:

$$\boldsymbol{H}_i=\text{Transformer}\left(\boldsymbol{Y}_i,E,\sum_{i=1}^{N}\text{sim}(\boldsymbol{Y}_i,\boldsymbol{E}_i)\right) \tag{28}$$

F. Emotion Classification and Model Optimization

During the emotion prediction stage, the composite vector is passed through a fully connected layer for affective classification. The ReLU activation function and the softmax function are used to project the affective features into the emotion category space, producing the classification output:

$$\boldsymbol{H}_i^{'}=\text{ReLU}(\boldsymbol{W}_u\boldsymbol{H}_i+\boldsymbol{b}_1) \tag{29}$$

$$Q_i=\text{Softmax}(\boldsymbol{W}_u\boldsymbol{H}_i^{'}+\boldsymbol{b}_2) \tag{30}$$

where, W_u denotes the weight matrix of the fully connected layer, b_1 and b_2 are bias vectors, and Q_i represents the softmax output. The final predicted emotion category is determined as $\hat{y}_i$. To optimize all parameters in the model, the minimized standard cross-entropy loss function is employed.

$$\hat{y}_i=\text{argmax}(Q_i) \tag{31}$$

$$L(\theta)=-\sum_{i=1}^{N}\sum_{j=1}^{c(i)}\log P_{i,j}\left[y_{i,j}\right] \tag{32}$$

IV. Empirical Results

A. Experimental Environment

The experimental environment was configured using the Ubuntu 22.04 operating system. Python 3.10 was adopted as the programming language, with Anaconda 4.2.0 used for package and environment management. PyTorch 1.12.1 served as the deep learning framework, supported by a 10 GB GPU. Computational efficiency and stability were ensured through the integration of CUDA Toolkit 11.3 and cuDNN v8.3.2 acceleration libraries.

B. Dataset and Evaluation Metrics

1) Self-constructed multimodal dataset: The dataset used in the experimental process was collected from students of the School of E-Commerce at the host institution. The intelligent interaction data were obtained from a representative application case titled AI Heart World, a typical intelligent non-cognitive ability assessment system based on immersive interaction. The system was developed around the core elements of "assessment needs – assessment environment – analytical techniques – information fusion" and integrates sandbox theory, psychometrics, game-based interaction, and artificial intelligence technologies. This interactive environment captures multimodal emotional information generated through scenario-based game interactions, enabling the measurement of an examinee's psychological state, personality traits, and behavioral characteristics in a more naturalistic setting, thus yielding more authentic data [22]. The collected data were processed and organized into a multimodal psychological assessment database (EC-SFED), which provides the foundation for Transformer-based sentiment analysis to effectively represent individual psychological states.



Fig. 3. Selected examples from the EC-SFED dataset.

During the experimental process, audio–visual information was collected from 82 university students using their personal devices. Each sample consisted of text, audio, and video components. Textual data were obtained from students' subjective responses within the intelligent interaction platform, including discussion content, forum exchanges, and instant messaging records. Audio data were recorded through microphones or headset devices during the psychological assessment sessions and captured acoustic characteristics such as intonation, loudness, and speaking rate. Video data were captured in real time via intelligent cameras, recording facial expressions, body posture, gaze dynamics, and other visual cues relevant to emotional analysis. The final dataset contained 720 text entries, 233 dialogues, 408 sentences, and 345 video clips. Each video clip was annotated with an emotion intensity score ranging from –3 to +3 (where –3 indicates highly negative emotion and +3 indicates highly positive emotion). Dialogues involved two or more speakers and were divided into training, validation, and test sets using an 8:1:1 ratio. Fig. 3 provides sample pictures of the participants. For feature extraction, the RoBERTa-base model was employed to generate 1,024-dimensional textual representations. Audio features were extracted using the Waveform-to-Vector 2.0 (Wav2Vec2) pretrained model, yielding 768-dimensional embeddings. Visual features were extracted using the ViT model, resulting in 342-dimensional representations. Multiple rounds of pretraining were conducted, and the final reported performance corresponds to the averaged results of 10 independent random runs on the test set.

2) Publicly available dataset: The CMU-MOSI dataset is a multimodal sentiment analysis dataset comprising three modalities of data - text, audio, and video-extracted from 93 YouTube movie review videos [23]. The text data is generated

through automatic transcription of the videos, while the audio and video data are directly extracted from the original footage. This dataset consists of a total of 2,199 video clips, each annotated with a sentiment intensity score ranging from -3 to 3. These annotations were completed by Amazon Mechanical Turk workers, who took comprehensive multimodal information into consideration.

Based on established findings in psychological assessment research, emotion prediction in this context is typically treated as a regression-oriented task. In the present study, the F1 score (W-F1) and classification accuracy (ACC) were adopted as the primary performance indicators. W-F1 reflects the model's overall classification capability, while ACC provides an intuitive measure of prediction correctness. The evaluation metrics are defined as follows:

$$\text{W-F1}=\frac{\sum_{a=1}^{R} M_a \times F1_a}{\sum_{a=1}^{R} M_a},\ \text{ACC}=\frac{\sum_{a=1}^{R} M_a \times ACC_a}{\sum_{a=1}^{R} M_a} \quad (33)$$

where, R denotes the total number of emotion categories in the dataset, M_a represents the number of samples belonging to category a, $F1_a$ is the F1 score for category a, and ACC_a is the accuracy score for category a.

C. Baseline Models

- MulT [7]: A bidirectional cross-modal attention mechanism is employed, using a modality-translation strategy to construct directed interactions between paired modalities. This design enables effective learning of cross-modal associative features.
- MISA [8]: A dual-space representation learning framework is introduced, in which distributional similarity, orthogonality constraints, reconstruction loss, and task-specific objectives are jointly optimized. The framework enhances cross-modal consistency while preserving modality-specific characteristics.
- CapTrBERT (Caption Transformer BERT) [17]: Semantic understanding is strengthened by generating auxiliary caption-like sentences and integrating them with the original textual input. This approach enables more comprehensive extraction of global semantic information.
- FITE (Face-sensitive Image-to-Emotional-Text Translation) [24]: A face-expression–to-text transformation strategy is adopted, in which facial expression features are converted into descriptive textual representations. These representations are aligned with target aspects at a fine-grained level, and multimodal information is fused through a gating mechanism for emotion classification.
- MIM (Multimodal Interaction Model) [25]: An interactive learning framework based on cross-modal attention is proposed. Modality-specific subspaces and cross-modal information exchange channels are constructed to facilitate efficient multimodal interaction while reducing feature redundancy.
- TIEMFF (Text Information Enhancement and Multimodal Feature Fusion) [26]: A text-dominant hierarchical enhancement architecture is introduced. A dual attention mechanism is used to strengthen textual features and achieve fine-grained alignment of multimodal representations.

D. Comparative Analysis of Experimental Results

Most of the video segments in the dataset were captured using high-definition devices, resulting in favorable lighting conditions and high visual clarity. Fig. 4 presents a comparison of evaluation metrics between the proposed model and the baseline models, including F1 scores and accuracy.

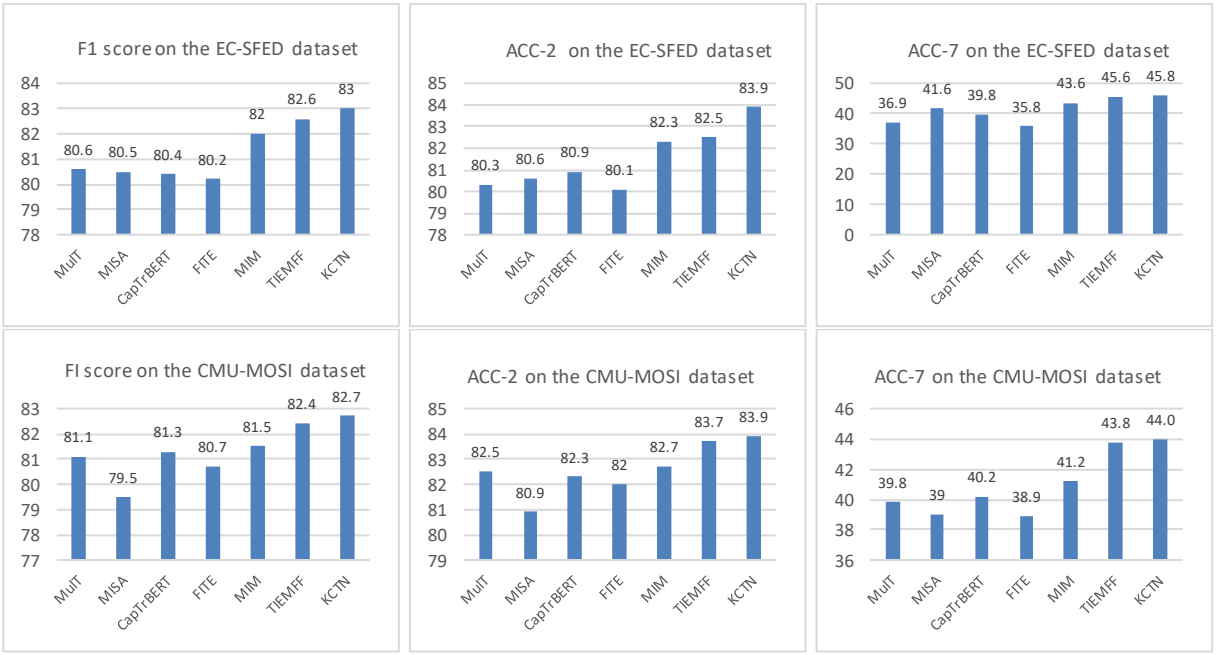


Fig. 4. Evaluation metrics of the proposed model and baseline models on the EC-SFED and the CMU-MOSI datasets.

The comparative analysis of experimental results obtained from the student psychological state dataset indicates that the performance of the FITE model is relatively weak, with both its F1 score and accuracy markedly lower than those of the other models. FITE transforms facial expression features into textual descriptions and subsequently performs multimodal fusion and emotion classification through a gating mechanism. In contrast, the MulT, MISA, and CapTrBERT models demonstrate moderate performance levels across the dataset, with comparable F1 scores and accuracy values. More recent models such as MIM and TIEMFF exhibit notably stronger performance, further validating the effectiveness of cross-modal attention mechanisms in sentiment analysis. The proposed KCTN model achieves the best performance across all evaluation metrics. The results confirm the model's effectiveness in cross-modal interaction and affective feature extraction while highlighting its robustness advantages in student psychological testing across diverse scenarios.

E. Ablation Study

1) Modality ablation: The influence of different modality combinations on the proposed model was examined. Performance was first evaluated using single-modality inputs—text (T), audio (A), and video (V). Dual-modality combinations, including text–audio (T–A), text–video (T–V), and audio–video (A–V), were then assessed. As expected, the ablation results obtained on the EC-SFED and the CMU-MOSI datasets are presented in Table II.

TABLE II. MODALITY ABLATION RESULTS ON THE EC-SFED DATASET

Modality combination	EC-SFED dataset		
ComParE 2016	F1 (%)	ACC-2 (%)	ACC-7 (%)
GeMAPS	80.6	80.3	46.8
	69.2	66.7	37.8
	72.1	71.5	39.8
	82.6	82.7	50.9
	83.3	83.2	51.9
	73.4	72.5	45.6
eGeMAPS	83.8	83.9	52.5

The results demonstrate that the performance of the multimodal affective recognition system is highly sensitive to the choice of modality combinations. In the single-modality experiments, notable disparities were observed across modalities: textual features achieved markedly superior performance (F1 = 80.6%) compared with audio (69.2%) and video (72.1%). The introduction of textual information produced substantial performance gains in dual-modality settings. The text–audio combination increased F1 by 19.36 percentage points relative to audio alone, while the text–video combination improved F1 by 15.53 percentage points. These results confirm the pivotal role of the textual modality in psychological assessment tasks, as its representational capacity surpasses that of auditory and visual modalities. Furthermore, single-modality comparisons revealed that audio conveyed more discriminative information than video, indicating that video data may contain higher levels of noise. The multimodal fusion experiments further demonstrate that combining text, audio, and video yields the highest overall performance. The tri-modal configuration achieved an F1 score of 83.8% on the EC-SFED dataset and 84.1% on the CMU-MOSI dataset, providing strong empirical support for the complementary nature of multimodal integration in psychological assessment scenarios.

2) Module ablation: To examine the contribution and influence of each core module within the overall architecture, a series of module-level ablation experiments was conducted on the two datasets. The notation "w/o" indicates the removal of a specific module during evaluation. Four key modules in the model were ablated sequentially, including the audio–visual enhanced Transformer module (w/o AVET), the information compression and fusion module (w/o ICF), and the long-range emotion fusion module (w/o FFM), enabling analysis of their respective impacts on model performance.

a) Ablation of the audio–visual enhanced transformer module (w/o AVET): In this setting, the audio–visual enhanced Transformer module of the CMET architecture was removed. The non-textual modalities processed by the one-dimensional temporal convolution were fed directly into the information compression and fusion module.

b) Ablation of the information compression and fusion module (w/o ICF): In this configuration, the information compression and fusion module of the CMET architecture was omitted. The modality representations produced by the non-textual enhanced Transformer were fused directly.

3) Ablation of the long-range emotion fusion module (w/o FFM): For this setting, the long-range emotion fusion module was removed. The audio–visual representations obtained from the information compression and fusion module were directly fused with the temporally convolved textual representations.

Table III presents the results of the module ablation experiments conducted on the EC-SFED and the CMU-MOSI dataset.

TABLE III. MODALITY ABLATION RESULTS ON THE EC-SFED AND THE CMU-MOSI DATASET

Model	EC-SFED dataset			CMU-MOSI dataset		
ComParE 2016	F1 (%)	ACC-2 (%)	ACC-7 (%)	F1 (%)	ACC-2 (%)	ACC-7 (%)
GeMAPS	81.3	82.5	50.5	78.0	78.3	47.8
	82.9	82.6	51.4	78.5	78.6	48.2
	82.2	83.3	52.6	80.3	81.2	48.5
	83.8	84.1	53.7	80.8	82.0	50.2

The module ablation results on the EC-SFED dataset demonstrate that the removal of any core module leads to a measurable decline in model performance. Among the examined components, the audio–visual enhanced Transformer module exerts the most substantial influence on overall effectiveness. Its removal reduces the F1 score to 81.3%, decreases ACC-2 to 82.5%, and lowers ACC-7 to 50.5%. This pronounced degradation underscores the central role of the

audio–visual enhanced Transformer in the early stages of multimodal information integration. The removal of the information compression and fusion module and the long-range emotion fusion module also results in significant performance degradation. These reductions highlight their importance in mitigating redundancy within non-textual modalities, enhancing feature robustness, and improving cross-modal discriminability. Ablation experiments conducted on the CMU-MOSI dataset also demonstrate the utility of the three aforementioned modules. Compared with the experimental results on the EC-SFED dataset, removing the long-distance emotion fusion module on the CMU-MOSI dataset leads to a relatively smaller impact on model performance. This suggests that the role of long-range information may be limited on smaller-scale datasets. Conversely, on large-scale datasets, this module can significantly enhance the model's ability to recognize both consistency and divergence in emotional information across different modalities by constructing a greater number of sample pairs, thereby improving the accuracy and effectiveness of sentiment analysis. Collectively, the ablation of any single module results in performance degradation, validating the contribution of each component to the model's overall performance and robustness.

F. Empirical Results and Analysis

To further evaluate the effectiveness of the proposed model in psychological testing, t-SNE was employed to visualize the high-dimensional features extracted from the EC-SFED test set. The results are presented in Fig. 5 and 6. Fig. 5 illustrates the distribution of feature representations before classification. Substantial overlap is observed between different emotional categories, and no clear clustering structure is formed. This indicates that the raw features do not inherently exhibit strong discriminative capability. After processing by the proposed model, the post-classification visualization in Fig. 6 shows that the feature vectors are distinctly clustered into two separable groups corresponding to positive and negative emotional states. The boundary between the clusters becomes pronounced, demonstrating that the model achieves effective multimodal affective feature separation. These empirical findings indicate that the proposed sentiment analysis model exhibits reliable performance when applied to psychological assessment scenarios involving students.

G. Case Analysis of Intelligent Psychological Assessment Interactions

The tasks in AI Heart World are presented within dynamic and immersive interactive scenarios. Participants' responses—including every choice made and action taken within these environments—generate process data with significant reference value for psychological state assessment. Table IV summarizes three representative cases, illustrating how different models handle these multimodal cues. In Case 1, the critical challenge lies in aligning textual information with visual cues from the scenario. The CapTrBERT model combines original inputs with generated auxiliary sentences; however, its architecture does not incorporate an effective mechanism for extracting or aligning emotional features from images. As a result, errors arise in sub-scenarios where textual expressions and visual affective cues are inconsistent. The FITE model establishes an initial degree of text–image integration through its basic cross-modal attention mechanism. In contrast, the proposed model leverages multimodal information captured throughout the intelligent interaction process to accurately identify the test subject's underlying anxious emotional state, which remains implicit beneath a superficially calm textual expression.

Case 2 focuses on a sub-scenario in which the visual information is incomplete, thereby assessing the model's ability to utilize non-facial emotional cues. In this scenario, the participant's face is partially occluded, resulting in the absence of clear facial features. Because the core design logic of the FITE model relies heavily on facial expressions as the primary basis for emotion inference, the absence of key visual features prevents the model from extracting effective supplementary cues from other regions of the image, ultimately leading to a misclassification of the psychological state. In contrast, the proposed model maintains prediction accuracy by leveraging the coordinated effects of the long-range contextual module and the Graph Convolutional Network (GCN), which together enable the extraction of syntactic features.

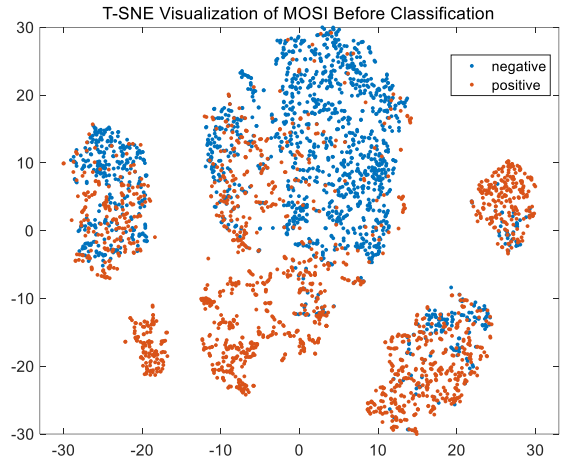


Fig. 5. Visualization of feature vectors before classification.

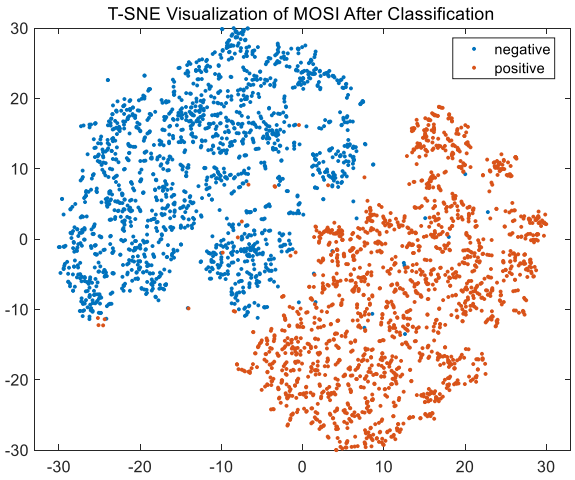


Fig. 6. Visualization of feature vectors after classification.

TABLE IV. CASE ANALYSIS OF PSYCHOLOGICAL ASSESSMENT SCENARIOS

Image				
Feature text		Did I do something wrong? Could you give me a hint?	I don't have enough blocks. They may not be stable.	The whole process was like playing a game. Using the sandbox to create scenes felt very novel, and it was quite interesting how the system analyzed my psychological state based on the 'Bing Dwen Dwen' character I selected.
Long-range text		This is impossible! I don't even know how to assemble the basic structure—forget it, let's move to the next one.	Try to speed up the construction. Use whatever is available first—it'll be finished soon.	Haha, this is kind of fun! Let me try another construction method!
Ground-truth emotion		Negative	Positive	Positive
Model	**FITE**	Negative√	Negative×	Positive√
	CapTrBERT	Positive×	Positive√	Positive√
	Proposed	Negative√	Positive√	Positive√

Case 3 represents an ideal scenario in which multimodal information is highly consistent, designed to validate the effectiveness and rationality of the evaluation benchmark. In this case, the participant's textual response (e.g., "It's quite interesting"), visual indicators (slight smile and relaxed posture), and contextual emotional records exhibit strong coherence, with no conflicting affective cues. Experimental results show that CapTrBERT, FITE, and the proposed model all achieve 100% prediction accuracy in this ideal setting. This outcome not only confirms that the baseline performance of each model meets expectations when sufficient information is provided, but more importantly, it validates the scientific rigor of the selected case set used in this evaluation. By employing a consistent scenario to eliminate external interference, the study successfully pinpoints the genuine performance differences among models under complex conditions, thereby establishing a reliable benchmark for subsequent model optimization.

V. CONCLUSION

A cross-modal fusion network model (KCTN) based on long-range affective integration was proposed for psychological assessment tasks involving university students. Utilizing the "AI Heart World" application case, the model intelligently extracted key information and incorporated a multimodal fusion module enhanced with long-distance knowledge, achieving layer-by-layer integration of textual, audio, and visual features. This approach effectively mined valuable information from weak modalities while suppressing noise interference. In practical psychological assessment tasks, KCTN was capable of handling imbalanced data quality in multimodal inputs. For instance, when audio or video data were partially missing or exhibited a low signal-to-noise ratio, the model maintained stable performance by integrating long-distance knowledge with the textual modality. In complex scenarios such as teacher-student dialogues or psychological counseling sessions, KCTN dynamically adjusted predictions based on contextual and emotional cues of the speaker, thereby improving recognition accuracy. Future research will continue to explore the disparities and complex interactions among different modalities to design more effective multimodal fusion mechanisms. It will also further integrate campus management systems and dialogue information from intelligent interactive platforms to uncover the causal layers underlying students' psychological states, thereby enabling timely detection of adverse mental conditions and contributing to the development of a novel intelligent technology-based paradigm for psychological assessment.

ACKNOWLEDGMENT

This study was funded by Zhejiang Provincial Education Science Planning Project (Research on the Construction of Multi modal Curriculum Resources and Personalized Teaching in Vocational Colleges Empowered by Embodied Intelligence, Grant No.: 2025SCG340); and Zhejiang Provincial Department of Education Project (Research on the Development and Validation of an Intelligent Monitoring System for College Students' Mental Health Integrating DT-IRTI Technology, Grant No.: Y202456020).

REFERENCES

[1] K. Wang, Y. Zhang, "Topic sentiment analysis in online learning community from college students," J. Data Inf. Sci., vol. 2, pp. 33–61, 2020.

[2] A. Iram, "Sentiment analysis of student's facebook posts," in *Int. Conf. Intell. Technol. Appl.*, pp. 86–97, 2018.

[3] M. Zhang, S. Wang, K. Yuan, "Sentiment analysis of barrage text based on ALBERT and multi-channel capsule network," in *Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.*, pp. 718–726, 2021.

[4] Y. F. Song, G. Ren, Y. Yang, X. C. Fan, "Multimodal sentiment analysis based on hybrid feature fusion of multi-level attention mechanism and multi-task learning," Appl. Res. Comput., vol. 39, no. 3, pp. 716–720, 2022.

[5] L. Sun, Z. Lian, B. Liu, J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," IEEE Trans. Affect. Comput., vol. 15, no. 1, pp. 309–325, 2023.

[6] R. Huan, G. Zhong, P. Chen, R. Liang, "Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences," IEEE Trans. Multimed., vol. 26, pp. 5753–5768, 2023.

[7] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, "Multimodal transformer for unaligned multimodal

language sequences," in *Proc. Assoc. Comput. Linguist.*, 2019, pp. 6558–6581.

[8] D. Hazarika, R. Zimmermann, S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimed.*, Seattle, WA, USA, pp. 1122–1131, 2020.

[9] W. Han, H. Chen, S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," arXiv preprint arXiv:2109.00412, 2021.

[10] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM," Multimed. Tools Appl., vol. 80, no. 9, pp. 13059–13076, 2021.

[11] J. H. Luo, Y. Zhu, "Multi interaction perception network for sentiment analysis of misaligned multimodal language sequences," Comput. Appl., vol. 44, no. 1, pp. 79–85, 2024.

[12] L. Yang, J. H. Zhong, Y. Zhang, X. Y. Song, "Temporal multimodal sentiment analysis with composite cross modal interaction network," Comput. Sci. Explor., vol. 18, no. 5, pp. 1318–1327, 2024.

[13] P. Zhong, D. Wang, C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," arXiv preprint arXiv:1909.10681, 2019.

[14] H. B. U. Haq, W. Akram, M. N. Irshad, A. Kosar, M. Abid, "Enhanced Real-Time Facial Expression Recognition Using Deep Learning," Acadlore Trans. AI Mach. Learn., vol. 3, no. 1, pp. 24–35, 2024.

[15] S. Nadar, D. Gandhi, A. Jawale, S. Pawar, R. Prabhu, "Multimodal Audio Violence Detection: Fusion of Acoustic Signals and Semantics," Acadlore Trans. AI Mach. Learn., vol. 4, no. 4, pp. 301–311, 2025.

[16] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," arXiv preprint arXiv:1906.05317, 2019.

[17] Z. Khan, Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. 29th ACM Int. Conf. Multimed.*, pp. 3034–3042, 2021.

[18] F. Eyben, M. Wöllmer, B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimed.*, Firenze, Italy, pp. 1459–1462, 2010.

[19] Y. Zhang, *Research on emotion analysis based on cross modal enhanced transformer fusion network*, Chang'an Univ., 2024.

[20] J. Chen, H. Hou, J. Gao, Y. Ji, T. Bai, "RGCN: Recurrent graph convolutional networks for target-dependent sentiment analysis," in *Int. Conf. Knowl. Sci. Eng. Manag.*, pp. 667–675, 2019.

[21] J. Jiang, A. Wang, A. Aizawa, "Attention-based relational graph convolutional network for target-oriented opinion words extraction," in *Proc. 16th Conf. Eur. Chap. Assoc. Comput. Linguist. Main Vol.*, pp. 1986–1997, 2021.

[22] K. Q. Huang, Y. X. Kang, C. X. Yan, et al., "A review of intelligent psychological assessment based on interactive environment," Chin. J. Ment. Health, vol. 39, no. 4, pp. 337–343, 2025.

[23] A. Zadeh, R. Zellers, E. Pincus, L. P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," arXiv preprint arXiv:1606.06259, 2016.

[24] H. Yang, Y. Zhao, B. Qin, "Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis," in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.*, pp. 3324–3335, 2022.

[25] Y. Luo, R. Wu, J. Liu, X. Tang, "Balanced sentimental information via multimodal interaction model," Multimed. Syst., vol. 30, no. 1, p. 10, 2024.

[26] Z. Liu, L. Cai, W. Yang, J. Liu, "Sentiment analysis based on text information enhancement and multimodal feature fusion," Pattern Recognit., vol. 156, p. 110847, 2024.

Enhancing Traffic Congestion Forecasting with Explainable Deep Learning: A Framework Using LIME for Transparent Intelligent Transportation Systems

Ouhmidou Hajar, Nabou Abdellah, Elikram Moulay Ahmed
Department of Computer Science-Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco

***Abstract*—Intelligent transportation systems aim to improve traffic management and road safety, manage traffic effectively, and reduce roadway system congestion. This optimally requires estimating future traffic congestion. Unfortunately, the most popular machine learning and deep learning techniques can be unsuitable for model development in this task due to interpretability challenges. This study attempts to provide a solution to this challenge by creating a tool that integrates Local Interpretable Model-agnostic Explanations (LIME) into any traffic congestion forecasting system. This tool is applied to the Metro Interstate Traffic Volume dataset, which contains samples of traffic and road system congestion along with temporal, weather, and contextual data. For global feature analysis, a Random Forest Regressor is used as a baseline model, while a neural network model is developed to predict the congestion of the traffic and road system. The neural network model achieved a congestion prediction with an R^2 score of 0.612, a mean squared error of 0.026, and a mean absolute error of 0.129. The LIME tool also provides temporal feature insights, which show that examples of weekday/holiday status reduce the sample congestion prediction for the example, while precipitation increases it. At a global level, hour of the day, day of the week, temperature, and month of the year are the dominant factors in congestion prediction. These findings illustrate the value of adding interpretability to predictive models of traffic congestion when using explainable artificial intelligence.**

***Keywords*—Explainable artificial intelligence; lime; traffic congestion forecasting; intelligent transportation systems; random forest**

I. Introduction

Accelerating urbanization and rapid population growth have presented problems for global transport infrastructure. Rapid increases in the number of cars have caused traffic congestion to be a day-to-day experience for most large cities, generating many effects ranging from economic loss of fuel to deterioration of air quality related to human well-being to lowered quality of life. To address those challenges, accurate and real-time traffic estimation was a primary objective of both researchers and transport agencies. These estimations are foundational to intelligent transport systems (ITS) for improving traffic, reducing travel time, and improving road safety [2]. Traditionally, historical and econometric models based on ARIMA and regression analyses have been dominant for traffic forecasting.

Although the output was extremely relevant for acquiring deep insight into traffic improvement, these modelling approaches, quite expectedly, do not lend themselves well to representing the complex non-linearities found in traffic because of heterogeneous and dynamically fluctuating factors such as climate and large events (festivals) and driver behaviour [3]. Recent advancements in deep learning (DL) have significantly transformed the traffic forecasting arena. The power of artificial neural models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks is demonstrated in their modelling of spatio-temporal dependencies and revealing patterns hidden in the raw traffic dataset [4]. Artificial neural models have consistently achieved state-of-the-art performance across an exceptionally wide range of applications that involve real-world processes [1]. However, along with the potential of improved accuracy brought by Deep Learning (DL) models comes the cost of non-interpretability; decisions made on top of complicated tower stacks of nonlinear transformations are opaque, and therefore DL models have also been termed "black box" models. In some applications, such as traffic management, where important decisions are made, the lack of interpretation can deter use cases because decision makers require both predictive validity and an explanation that they can trust. For example, traffic operators need to understand why a model predicted significant congestion at a certain time and location, validate the analysis, the reason for the analysis, and take appropriate action in time. To address this gap, the emerging field of Explainable Artificial Intelligence (XAI) has proposed different methods for explaining black-box models. Prominent methods such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive ExPlanations (SHAP) deliver post-hoc explanations for models in terms of how much the most relevant features contribute to single predictions [5],[6]. These methods provide meaningful information about model reasoning, supporting a better understanding and trust of the end user in the model's output. This revised manuscript addresses the methodological inconsistency identified during review by reframing the work as a model-agnostic explainability framework rather than as a purely deep-learning study. The predictive component may be implemented using either an ensemble model or a neural network, while LIME is used as a post-hoc explanation layer.

The main contributions are as follows:

- Evaluation framework: Propose a methodological framework for developing and evaluating interpretable traffic prediction models.
- Practical demonstration with LIME: Using a public dataset, apply LIME to explain the predictions of a congestion model and illustrate how local explanations can enhance understanding.
- Analysis of influential features: Identify and visualize both local (LIME-based) and global (feature importance-based) drivers of congestion predictions.
- Reproducible resources: Provide fully reproducible Python code to support researchers and practitioners in adapting and extending the approach.

The rest of this study is structured as follows. Section II presents related works on traffic prediction and interpretability. Section III explains the approach, starting from the description of the dataset preprocessing and predictive model, followed by LIME integration. Experimental results, visual analysis, and comparative studies are reported in Section IV, while the study is finally concluded in Section V, which describes future research directions.

II. Related Work

The field of traffic forecasting has undergone a significant paradigm shift over the last decade. Early methodologies primarily relied on classical statistical frameworks and econometric models, such as ARIMA and Kalman filters. While these approaches provided foundational insights into traffic patterns, they often struggled to account for the volatile, non-linear dependencies caused by heterogeneous factors like sudden weather shifts or large-scale urban events [7]. With the rise of "Big Data," researchers transitioned toward machine learning techniques, including Support Vector Machines (SVM) and Random Forests, to better capture these complexities. However, the most transformative leap occurred with the introduction of deep learning (DL). Architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) [8] networks have demonstrated an unparalleled ability to model spatio-temporal dependencies, consistently achieving state-of-the-art performance across diverse real-world datasets. Despite these gains in predictive power, the "black-box" nature of deep neural networks remains a critical bottleneck. In the context of Intelligent Transportation Systems (ITS), raw accuracy is insufficient if the underlying logic is opaque. Traffic operators [9] require "trustworthy" systems that allow them to validate predictions and perform root-cause analysis before committing to high-stakes management decisions. To bridge this accountability gap, the emerging field of Explainable Artificial Intelligence (XAI) has introduced post-hoc interpretability tools [10]. Methods such as Local Interpretable Model-agnostic Explanations (LIME) have gained prominence for their ability to provide human-understandable justifications for individual model decisions without sacrificing the performance of the underlying complex model. This research builds upon this momentum, proposing a unified framework that integrates LIME with deep learning to provide the transparency required for modern, resilient urban mobility.

For traffic and mobility applications, explainability can reveal whether congestion forecasts are driven by time of day, weather, holiday status, or other contextual factors. Visual analytics approaches have also been developed to help domain experts understand congestion-influencing factors using explainable machine learning [15]. The present work contributes to this line of research by combining local LIME explanations with global feature importance in a transparent traffic-forecasting workflow.

III. Methodology

The predictive performance and interpretability of the traffic congestion prediction framework are emphasised. The system is built on four main components. The first is the dataset. The second is data preprocessing. The third is predictive modelling. The fourth is the use of LIME for interpretability. A reliable and reproducible approach to creating traffic prediction systems is provided by these components when used together.

A. Dataset

This study uses the Metro Interstate Traffic Volume dataset from the UCI Machine Learning Repository [12], which is publicly accessible. The data set contains hourly traffic volume counts for a section of the I-94 interstate highway in Minnesota, USA. It also includes context variable information to show factors affecting traffic flow.

Table I summarizes the main variables used from the Metro Interstate Traffic Volume dataset, including temporal, weather, contextual, and traffic-volume attributes.

TABLE I. The Main Dataset Variables

Variable	Description
Holiday	Indicator of public holiday status.
Temperature	Recorded in Kelvin.
Rain and Snow	Precipitation levels in millimetres
Clouds	Percentage cloud cover.
Date time	Hourly timestamp used to derive hour, day of week, month, and year.
Weather main and Weather description	General and detailed weather conditions.
Traffic volume	Target variable, representing the number of vehicles per hour.

The presence of meteorological, time, and contextual variables makes this dataset particularly well-suited for studying real-world traffic dynamics and prediction, which is a significant advantage.

B. Data Processing

The preprocessing pipeline was revised for clarity and reproducibility. First, the timestamp was converted to a datetime format and decomposed into hour, day of week, month, and year. Second, categorical variables such as holiday and weather conditions were encoded using one-hot encoding to avoid artificial ordinal relationships. Third, the target variable was separated from the explanatory variables. Fourth, the data were

divided into training and testing subsets using an 80/20 split with a fixed random seed. Finally, scaling was applied where required by the neural-network model and by the LIME prediction wrapper.

C. Prediction Model

To demonstrate interpretability, employ a Random Forest Regressor as the predictive model. Random Forests, an ensemble of decision trees, are robust against overfitting and capable of modelling complex nonlinear relationships. While not the most advanced option compared to deep neural networks, Random Forests serve as an effective and transparent example for integrating interpretability tools such as LIME. Importantly, because LIME is model-agnostic, the same interpretability process can be extended to deep learning architectures.

A neural network regressor is also evaluated on normalized congestion values. It is implemented as a feed-forward multilayer perception operating on the engineered tabular variables. The input layer corresponds to the number of preprocessed features, hidden dense layers learn non-linear interactions, and the output layer contains a single regression neuron. The network is trained using Mean Squared Error as the loss function and monitored through validation loss and Mean Absolute Error. This formulation is consistent with the reported training and validation curves and avoids presenting the Random Forest as a deep-learning model.

The two models are not used to claim state-of-the-art performance. Instead, they demonstrate that the same explanation framework can be connected to different predictive engines.

D. Deep Learning Network Architecture for Traffic Prediction

Deep learning has emerged as a transformative approach to traffic prediction. Models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are achieving state-of-the-art results. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), which are achieving state-of-the-art results. These architectures excel due to their ability to capture spatio-temporal dependencies on traffic data, i.e., patterns that evolve over time and space. Long Short-Term Memory (LSTM) networks are well-suited to time series forecasting. By retaining information over long sequences, LSTMs can model long-term dependencies in traffic flow, which is a capability that traditional methods often lack.

Fig. 1 presents the overall workflow adopted in this study, starting from traffic data collection and verification, followed by data curation, feature extraction, analytical processing, prediction, and validation. This framework highlights the complete pipeline used to transform raw traffic-related data into interpretable forecasting output.

A typical LSTM-based setup receives sequential traffic data and learns to pick up on repeated patterns and trends in the data, which are then used to predict future traffic. Unlike the old ways of coping with prohibitive amounts of feature engineering, these networks learn relevant representations from raw data automatically. Such an ability to unveil hidden patterns in large data renders deep neural networks excellent instruments for intelligent transport systems (ITS), where timely and precise traffic prediction is important to improve road safety, reduce congestion, and optimize urban mobility.

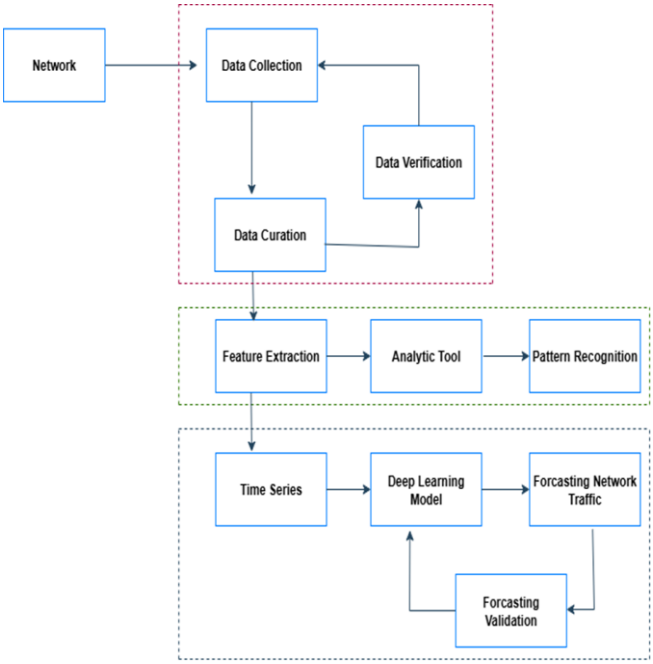


Fig. 1. End-to-end framework for traffic data collection, analysis, and prediction.

E. Lime Architecture

The primary idea behind LIME (Local Interpretable Model-agnostic Explanations) is to increase the transparency of complex models by creating interpretable, intuitive approximations for nearby individual predictions [11]. Unlike explaining the entire model globally, LIME addresses only a local neighborhood around a specific instance. The process begins by perturbing the input data point to create a set of nearby samples. The original model then makes predictions on the perturbed instances. From these predictions, LIME derives a sparse, interpretable model of ten times a linear regressor wherein the contribution of each sample is weighted by its proximity to the original instance.

F. Mathematical Problem Formulation

The primary objective of traffic congestion forecasting is to map a high-dimensional set of historical and contextual features to a continuous target variable representing future traffic volume. Formally, define the traffic state at a specific time interval t as $x_t \in R$ the task is to construct a predictive mapping function $f: x \rightarrow y$, where the input space $\mathcal{X}$ is composed of a multidimensional feature vector $\mathcal{V}_t$. This vector incorporates temporal attributes (hour, day, month), meteorological data (temperature, precipitation), and contextual indicators such as holiday status.

The forecasting model aims to minimize the discrepancy between the predicted volume $\hat{y}_t$ and the ground truth observation obtained from the Metro Interstate dataset. This is

achieved by optimizing a loss function, specifically the Mean Squared Error (MSE), defined as:

$$MSE = \frac{1}{n} \setminus \sum_{i=1} (y_i - \hat{y}i)^2$$

To ensure transparency, augment this predictive model with a local surrogate explanation model g . For a specific traffic instance x, the explanation $\xi(x)$ is derived by solving the following optimization problem:

$$\xi(x) = argmin_{g \in \mathbb{G}} \mathcal{L}(f, g, \pi x) + \Omega(g)$$

In this formulation, $L(f, g, \pi x)$ represents the fidelity loss, which measures how accurately the interpretable model g approximates the complex "black-box" model f within the local neighbourhood πx . The term $\Omega_{(g)}$denotes the complexity of the explanation, ensuring that the resulting insights remain human-interpretable. By minimizing this joint objective, the framework identifies the specific features such as heavy rain or peak commute hours that exert the most significant influence on a given congestion forecast [13].

The Fig. 2 Effect of LIME original prediction explained by fake data around the input instance fitted with a locally weighted interpretable model. While the example is shown for image prediction, applying this approach in the work to tabular traffic data can concisely explain individual congestion forecasts.

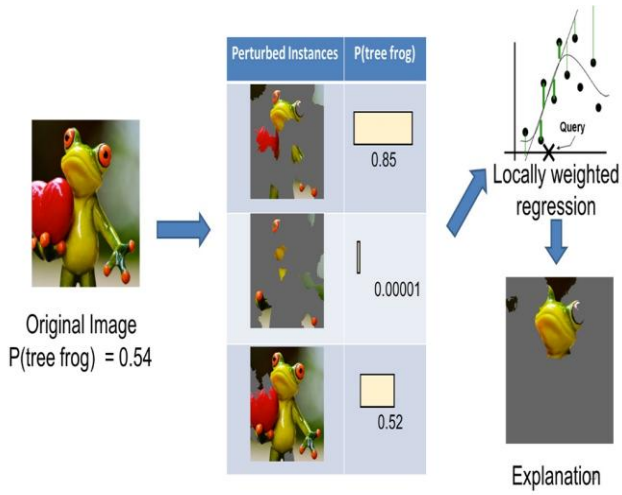


Fig. 2. Local Interpretable Model-Agnostic Explanations (LIME) applied to image predictions.

The coefficients of the surrogate model act as explanations, indicating which of the features contributed most to the prediction of the original model for that sample. In this way, LIME provides a localized, human-interpretable view of the reasoning of a "black box" model. Rather than giving an overall view, it illuminates decision-making in specific settings, giving practitioners valuable insight into why the model predicted the outcome in question.

G. Lime Integration

To address the "black box" nature of predictive models, integrate Local Interpretable Model-agnostic Explanations (LIME) into the framework. LIME provides local human-understandable explanations for individual predictions through the following procedure:

- Initialization: A LimeTabularExplainer instance is built using the training data, feature names, and the regression mode.
- Instance Selection: A representative instance from the test set is selected for explanation, allowing us to examine model reasoning for a specific traffic scenario.
- Prediction Function Wrapper: A custom predict fn raw function is defined. This wrapper takes raw feature inputs, scales them using the previously fitted StandardScaler, and then passes them to the trained DNN model for prediction.

This ensures that LIME interacts with the model in its expected scaled input space while generating explanations based on raw feature values.

IV. Explanation Generation

LIME perturbs the selected instance and evaluates the model's predictions on these variations. It then fits a simple interpretable model (e.g., a local linear model) to approximate the complex model's behaviour in the vicinity of the instance.

Through this process, LIME highlights the relative contribution of input features (e.g., weather, time of day, holidays) to a given prediction. These explanations make the model's reasoning more transparent, bridging the gap between predictive performance and interpretability in traffic management systems. The experimental evaluation highlights both the predictive performance of the model and the interpretability provided through LIME.

In addition to presenting standard performance metrics, include visual explanations that illustrate how different features contribute to traffic congestion predictions at both the local and global levels.

A. Model Evaluation

The revised evaluation separates raw-volume results from normalized-congestion results to avoid misleading comparisons. The Random Forest result is reported on raw traffic-volume units, whereas the neural-network result is reported on normalized congestion values. Therefore, the numerical values of the errors should not be compared directly unless the same target scale, preprocessing, and split are used.

TABLE II. Reported Model Performance

Model	Target Scale	Metrics reported	Interpretation
Random Forest Regressor	Raw traffic volume	MSE= 189,619.42	Baseline model and source of global feature importance.
Neural network regressor	Normalized congestion	R^2=0.612; MSE=0.026 MAE=0.129	Predictive model used for the reported training curves and local LIME explanations.

Table II presents the reported performance metrics for the Random Forest Regressor and the neural network regressor, while clarifying the target scale used for each model.

The neural network explains approximately 61.2% of the variance in normalized congestion values. Fig. 3 and 4 show that training and validation errors decrease over the training epochs, which indicates model convergence without a severe divergence between training and validation curves.

Fig. 3 shows the Training & Validation loss curves from model training. The gradual decrease in the curves indicates that the model learned the traffic pattern progressively. The training and the validation loss are rather similar (721 out of 884), indicating that the model converged well without any overfitting or very little evidence.

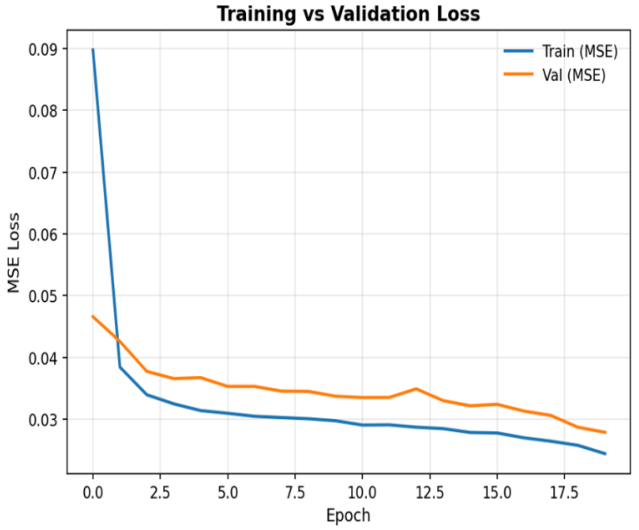


Fig. 3. Training and validation loss curves showing model convergence without overfitting.

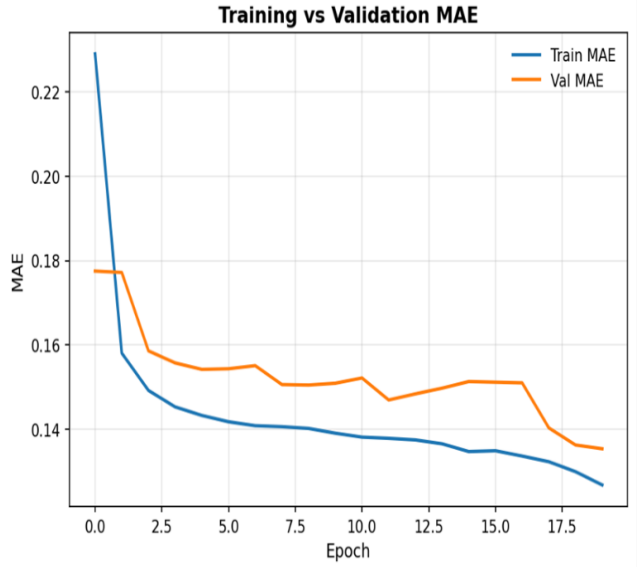


Fig. 4. Mean Absolute Error progression during training, demonstrating consistent improvement.

Fig. 4 shows the evolution of the Mean Absolute Error (MAE) during training and validation. The decreasing trend of both curves confirms that the prediction error was reduced over successive epochs. This behavior indicates that the model improved its forecasting capability as the training process progressed.

B. Prediction Accuracy Analysis

The predicted-versus-actual plot in Fig. 5 shows that predictions generally follow the diagonal reference direction, although dispersion remains visible. This confirms that the model captures a meaningful part of the congestion dynamics but also indicates that additional features, temporal lags, or more advanced architecture could improve prediction quality.

It's comparing the predicted congestion values to what was observed in the test set. The point distribution distributed around the diagonal reference line of the graph reveals that the model was able to reconstruct a general trend in traffic congestion. On the other hand, the dispersion of some points also indicates that due to those traffic conditions, there are corresponding prediction errors that require other interpretability tools to use, such as LIME.

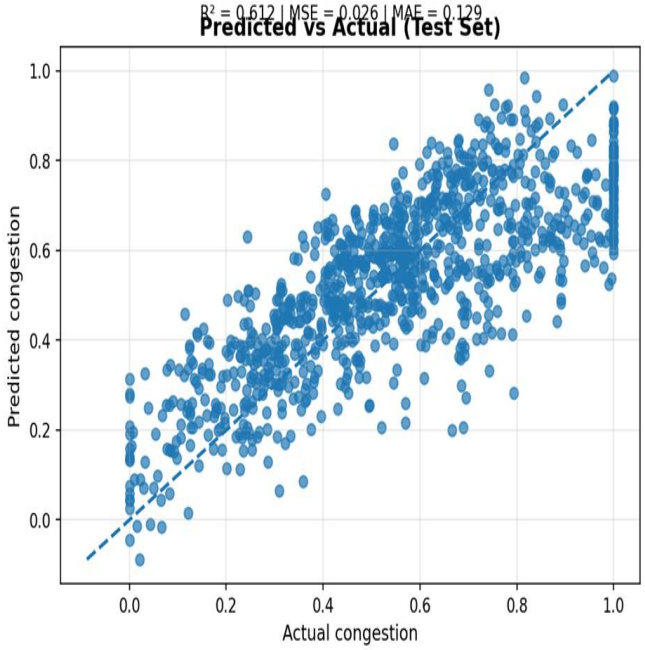


Fig. 5. Predicted vs. actual congestion values.

C. Residual Analysis

Residual diagnostics are presented in Fig. 6 and 7. The residual histogram is centered near zero, which suggests that the model does not show a strong average bias. The residuals versus predicted plot does not show a single linear systematic pattern, although the spread indicates that prediction errors vary across congestion levels. These diagnostics should be retained because they provide more information than aggregate metrics alone.

Fig. 6 shows the histogram of residuals (the gap between actual and predicted congestion values. Finding that the residuals mainly lie close to zero, indicating there is little global bias in the model. The nearly symmetric histogram also confirms that there are roughly as many overestimations as underestimations in the test set.

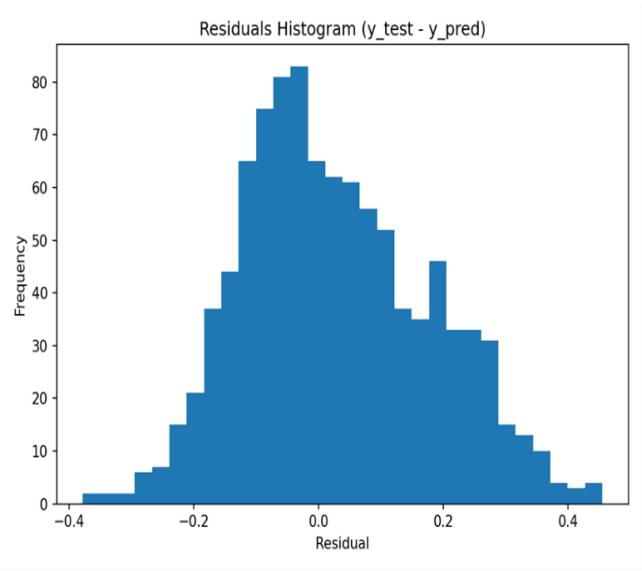


Fig. 6. Histogram of residuals showing approximately normal distribution centered at zero.

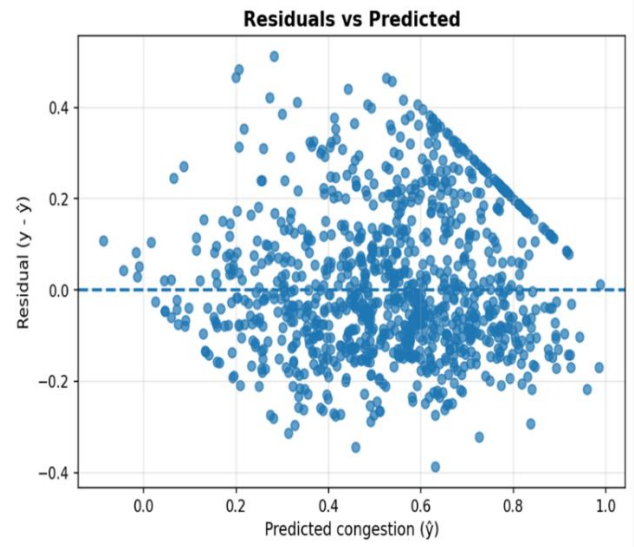


Fig. 7. Residuals vs. predicted values showing a homoscedastic pattern.

Fig. 7 shows the distribution of residuals with respect to the predicted congestion values. The absence of a clear systematic trend suggests that the model does not strongly overestimate or underestimate congestion within a specific prediction range. This supports the reliability of the prediction results, although some dispersion remains visible for medium and high predicted congestion values.

D. Local Explanations with LIME

To gain insight into individual predictions, LIME was applied to selected test instances. LIME produces bar plots that highlight the relative influence of features on a given prediction, with the bar's length and direction indicating the strength and sign of the contribution. Positive values increase the predicted traffic volume, while negative values decrease it.

LIME analysis provides crucial insights into feature importance for individual predictions:

Key Interpretability Findings:

- Day of Week (Dominant Factor): Shows the strongest negative contribution (-0.30), indicating weekends or specific days significantly reduce congestion prediction.
- Holiday Status: Secondary negative contributor (-0.15), confirming reduced traffic during holidays.
- Precipitation: Positive contribution (+0.08), suggesting increased congestion during rainy conditions.
- Temperature and Hour: Moderate influences on prediction, with temperature showing a slight negative impact and hour showing a minimal positive contribution.

Fig. 8 presents the local LIME explanation for the selected representative test instance. It shows which variables increase or decrease the predicted congestion value, thereby making the individual model prediction more transparent.

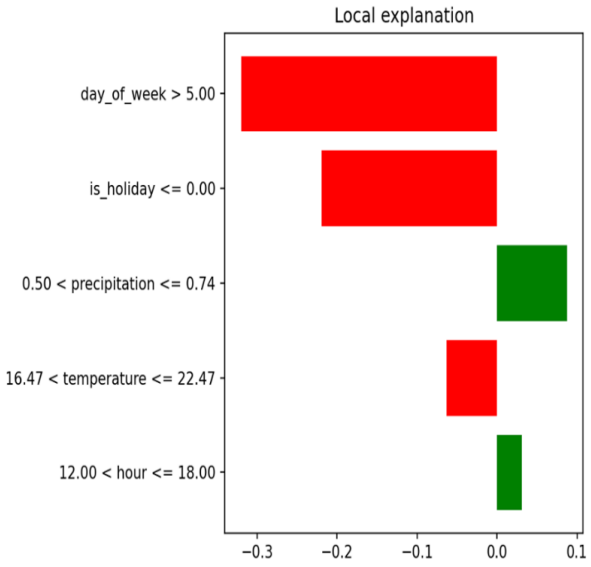


Fig. 8. LIME explanation for a representative test instance showing feature contributions.

Feature Interpretation;

- Red bars: Features decreasing congestion prediction
- Green bars: Features increasing congestion prediction
- Bar length: Magnitude of feature impact on this specific prediction

This LIME analysis demonstrates that temporal factors (day of week, holidays) are primary drivers of congestion prediction, while environmental factors (precipitation, temperature) provide secondary but meaningful contributions.

These local explanations make the model's decision-making process transparent, allowing practitioners to trace back a prediction to specific contributing factors.

Fig. 9 gives a more detailed local explanation for a single prediction with an estimated congestion value of 0.12. The result

confirms that LIME can identify the specific variables responsible for a given model output. In this instance, weekend-related temporal effects reduce the predicted congestion, whereas precipitation and afternoon timing slightly increase the forecasted value.

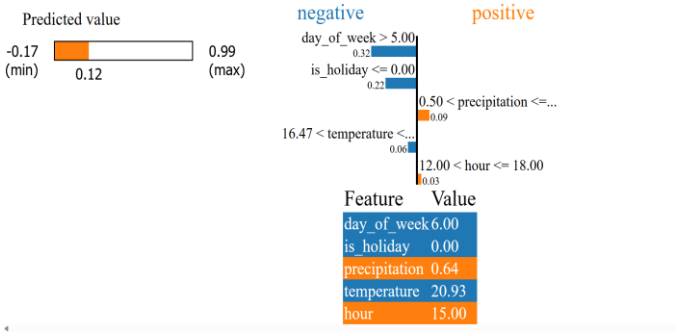


Fig. 9. LIME explanation of a single prediction (0.12), showing the local positive and negative contributions of temporal and weather-related features.

Analysis of the LIME explanation for a specific instance reveals a local prediction of 0.12, falling between a minimum value of -0.17 and a maximum value of 0.99. Factors with a significant negative impact on this prediction include the day of the week being a weekend (day 6.00, contributing negatively at 0.32) and the absence of a public holiday (isholiday 0.00, with a negative contribution of 0.22). A temperature above 16.47 (specifically 20.93) also has a slight negative effect (0.06). On the other hand, precipitation between 0.50 and a higher value (specifically 0.64) has a notable positive influence (0.09), as does the time of day between 12:00 p.m. and 6:00 p.m. (specifically 3:00 p.m.), with a positive contribution of 0.03. These results highlight the importance of weather conditions and timing in determining the model's prediction for this instance.

E. Global Feature Importance

Beyond individual predictions, it is also important to capture the global influence of features across the entire dataset. In Random Forests, feature importance is derived from the average reduction in variance contributed by each variable across all trees in the ensemble.

Fig. 10 presents the top 10 global feature importance rankings obtained from the Random Forest model. This global analysis complements the local LIME explanations by identifying the variables that most consistently influence congestion predictions across the dataset.

The top ten features identified included hour, temperature, and month, all of which consistently played a dominant role in shaping traffic patterns. This global feature importance ranking was visualized in a bar chart, showing which variables the model relies on most when generating predictions.

Together, the local explanations from LIME and the global feature importance analysis provide a complementary perspective. While LIME sheds light on why the model predicts a specific outcome in each context, feature importance summarizes the broader, dataset-wide drivers of congestion. This dual perspective not only improves transparency but also supports more informed decision-making in traffic management applications.

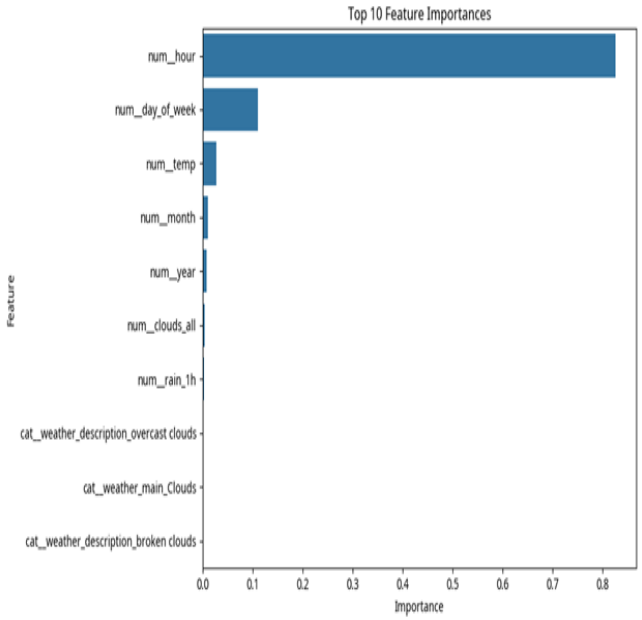


Fig. 10. Feature importance analysis for network traffic forecasting.

F. Positioning Against Related Work

Table III positions the results against representative studies. Direct numerical comparison is possible only when the same dataset, target scale, preprocessing pipeline, and temporal split are used.

TABLE III. COMPARISON WITH REPRESENTATIVE RELATED STUDIES

Study	Model or scope	Relation to this work
Lv et al. [7]	Deep learning for traffic-flow prediction using large-scale traffic data.	Supports the use of deep learning for non-linear traffic dynamics but uses a different dataset and setting.
Ma et al. [8]	LSTM model for traffic-speed prediction.	Shows the value of recurrent models for long-term temporal dependencies.
Polson and Sokolov [9]	Deep-learning architecture for short-term traffic-flow prediction.	Demonstrates the ability of deep models to capture traffic regime changes.
Li et al. [7]	DCRNN for graph-based traffic forecasting.	Addresses road-network spatio-temporal dependencies not modeled in the present single-station tabular setup.
Pranolo et al. [14]	LSTM and Bi-LSTM on the Metro Interstate dataset.	Provides a dataset-specific reference point; however, comparison requires identical preprocessing and splits.

G. Results and Discussion

The experimental results demonstrate that the deep learning framework effectively captures the complex dynamics of urban traffic, achieving an R^2 score of 0.612 and a Mean Squared Error (MSE) of 0.026. While these quantitative metrics confirm the model's predictive reliability, the integration of LIME provides a critical qualitative layer that moves beyond "black-box" forecasting. By decomposing individual predictions, identifying those temporal features, specifically the "Day of the Week," exerted the most significant influence on congestion levels, showing a strong negative contribution of -0.30. This suggests

that the model successfully learned the reduced commercial and commuter activity typical of weekends, allowing for more nuanced weekend-specific forecasts.

Furthermore, the model's sensitivity to environmental factors such as the positive contribution of precipitation (+0.08) indicates an accurate alignment with real-world traffic physics, where adverse weather typically reduces free-flow speeds and increases density. Unlike traditional global importance metrics, which provide a static view of the dataset, the local LIME explanations revealed that at specific peak hours, weather conditions can become the dominant driver of error if not properly accounted for. This dual-perspective analysis, combining global feature rankings with instance-specific explanations, equips traffic managers with the "reasoning" behind a forecast. Ultimately, this transparency transforms the AI from a mere prediction tool into a decision-support system that can be validated and trusted by human operators in real-time Intelligent Transportation Systems (ITS).

V. Conclusion and Future Work

This study successfully establishes a framework for integrating deep learning with Explainable AI (XAI) to resolve the "black-box" limitations of modern traffic forecasting. By developing a model that achieves a robust R^2 score of 0.612 demonstrates that high-fidelity predictive performance can coexist with technical transparency.

The integration of LIME provides a dual-layered interpretability that is essential for real-world deployment. On a local level, it empowers traffic managers to validate individual forecasts by identifying specific drivers such as the -0.30 impact of weekend temporal factors or the +0.08 influence of precipitation, allowing for targeted, data-driven interventions. Globally, the analysis confirms that variables such as the hour of the day and temperature remain the most consistent predictors of urban mobility patterns.

Ultimately, this research transitions AI from a mere forecasting tool into a trustworthy decision-support system. By providing the "reasoning" behind the numbers, this framework fosters the institutional trust required to implement safer, more efficient, and sustainable smart city infrastructures. Future work will focus on scaling these explanations to real-time dynamics and exploring the generalizability of this framework across more complex graph-based road networks.

Acknowledgment

The authors would like to thank the Faculty of Sciences Semlalia at Cadi Ayyad University for providing the research facilities and computational resources necessary to conduct this study. Also, thanks to Nabou Abdellah and Elikram Moulay Ahmed for their insightful feedback on the explainability framework and the application of LIME. Additionally, express gratitude to the anonymous reviewers for their constructive comments that improved the quality of this manuscript.

References

[1] M. Adnan and S. Islam, "A review on traffic prediction using machine learning and deep learning techniques," IEEE Access, 2019.

[2] Ribeiro, M. T. (n.d.). LIME - Local Interpretable Model-Agnostic Explanations. https://homes.cs.washington.edu/~marcotcr/blog/lime/

[3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017 *

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, 2015, doi: 10.1109/TITS.2014.2345663.

[5] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," Transportation Research Part C: Emerging Technologies, vol. 54, pp. 187–197, 2015, doi: 10.1016/j.trc.2015.03.014.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, 2015.

[8] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction," Transportation Research Part C: Emerging Technologies, vol. 54, pp. 187–197, 2015.

[9] N. G. Polson and V. O. Sokolov, "Deep learning for traffic flow prediction," Transportation Research Part C: Emerging Technologies, vol. 79, pp. 1–17, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[11] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[12] Hogue, J. (2019). Metro interstate traffic volume [Dataset]. In UC Irvine. https://doi.org/10.24432/c5x60b.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[14] A. Pranolo, S. Saifullah, A. B. U. Putra, R. Drzewski, and A. P. Wibawa, "Urban traffic volume prediction using LSTM and Bi-LSTM: Performance evaluation on the Metro Interstate dataset," ILKOM Jurnal Ilmiah, vol. 17, no. 3, pp. 227–240, 2025, doi: 10.33096/ilkom.v17i3.3001.227-240.

[15] X. Chen, J. Zhang, H. Wang, and X. Song, "TCEVis: Visual analytics of traffic congestion influencing factors based on explainable machine learning," Visual Informatics, vol. 8, no. 1, pp. 56–66, 2024, doi: 10.1016/j.visinf.2023.11.003.

Mammographic Image Classification Based on KAN-CBAM

Yuanyuan Wang[1], Vladimir Mariano[2]
College of Computing and Information Technologies, National University, Manila, Philippines[1, 2]
School of Information Engineering, Guizhou University of Traditional Medicine, Guizhou, China[1]

***Abstract*—With the fast enhancement of deep learning, research on automatic detection of breast tumors is becoming increasingly in-depth. However, traditional CNNs' linear kernel has difficulty not only in capturing the nonlinear combination of low-frequency structures and high-frequency details but also in fully exploring the nonlinear discriminative features in medical images. Furthermore, the current models used for breast tumor detection have complex structures and slow inference speeds. Therefore, this study solves the linear kernel problem and improves the model inference speed by using a lightweight mammographic image classification method based on KAN-CBAM to speed up breast cancer diagnosis. The proposed method introduces the KAN convolution module, which embeds a learnable B-spline activation function into the convolution kernel. This scheme improves the capability of the proposed method to capture nonlinear features and improves its capacity to fit complex, nonlinear distributions. Moreover, the proposed method combines the CBAM attention mechanism to screen key semantic channels through channel attention and then uses spatial attention to locate lesion areas, achieving "channel-space" dual feature recalibration, further improving the attention to key features, and achieving more accurate classification in complex and variable medical images. We evaluated the proposed method on the mammographic image datasets DDSM, INbreast, and MIAS to verify its performance. The results prove that KAN-CBAM models have higher adaptation to diverse dataset scales, efficiently acquiring major lesion parts and nonlinear discriminatory features in mammographic images. Meaningful and great enhancements were seen in different metrics such as accuracy, F1-score, AUC, precision, and recall, demonstrating extensively improved model strength and generalization capability.**

Keywords*—KAN; CBAM; deep learning; mammographic; classification

I. Introduction

A. Background of the Study

Breast cancer is one of the most common female fatal tumors characterized by abnormal proliferation of breast tissue and tumor formation. Its incidence rate exceeds even that of lung cancer, becoming the world's largest cancer [1][2]. Timely diagnosis of benign and malignant breast tumors allows for earlier treatment, thereby reducing patients' life risk and helping to decrease breast cancer mortality [3]. Therefore, the recognition of breast tumors is crucial to the judgment of breast cancer. Common breast imaging techniques include HP (Histo-Pathology), US (Ultrasound), MRI (Magnetic Resonance Imaging), DM (Digital Mammography), and mammography images [4], which serve important roles in the early identification of malignant lesions.

Currently, mammography is the only imaging modality widely recognized as suitable for large-scale screening of asymptomatic women and serves as the cornerstone of breast-cancer screening programs worldwide [5]. Because of its low cost and medical requirements, it has become one of the most widely used detection methods for breast cancer diagnosis [6]. In the context of breast cancer, a large number of researchers are committed to exploring strategies for classifying benign and malignant tumors [7]. In order to better detect and diagnose breast cancer cases on time, deep learning (DL) technology is a promising method for disease detection in the medical field. Compared with traditional image classification models such as Support Vector Machine (SVM) [8], Fuzzy Cognitive Map (FCM) [9], or Multi-Layer Perceptron (MLP) [10], deep learning techniques can automatically discover features in breast images that contain benign and malignant discrimination information, and have stronger breast feature extraction capabilities, such as convolutional neural networks (CNN) [11], which demonstrate powerful capabilities in image feature extraction. Based on CNN, various types of image classification models were constructed based on different learning algorithms [12]. Jafari, Z. et al. [13] proposed a CNN model based on feature extraction and dimensionality reduction for breast cancer detection in mammography images. According to the results, the method achieved better classification accuracy. Bouzar Benlab, L. et al. [14] proposed a new framework for breast cancer detection of breast X-ray images, called the CNN-CBR classification system. In order to improve the accuracy of CBR, the authors implemented image enhancement and data enhancement to optimize the effect of feature extraction. Efficiently obtaining regions of interest (ROI) in breast images through the U-Net architecture, this framework combines deep learning (DL) with CBR. Deep learning helps achieve accurate breast image segmentation, while CBR provides interpretable and accurate classification results. S. Majumdar et al. [15] combined three transfer CNN models, namely GoogleNet, VGG11, and MobileNetV3_Small, to solve the binary classification problem of breast tissue pathology images using the Gamma function. Moreover, J. G., T. Melekoodapattu et al. [16] designed an automatic diagnosis system for breast cancer. The system uses a CNN containing nine convolution layers to extract image texture features and then classify the extracted features. When extracting image-texture features, we reduce the dimension of breast-image representations layer by layer, thereby enhancing the expressiveness of the learned features and improving

breast-cancer diagnostic accuracy. In general, with CNN as the core, the automatic recognition accuracy and robustness of breast cancer images have been significantly improved through feature engineering, data enhancement, transfer learning, or fusion with traditional machine learning and other strategies.

Even though CNN-based image-feature techniques have achieved strong performance in medical imaging, the lesion regions in medical images exhibit highly complex characteristics. Existing neural networks based on fully connected layers still cannot solve the linear kernels problem, that is, the network has no perception in the spatial dimension and lacks nonlinear activation. The overall degradation is linear mapping, making it difficult to capture complex textures and deep semantics, and making it unable to effectively handle nonlinear classification problems of medical image data [17]. For this reason, the Kolmogorov-Arnold Network (KAN) has been applied to image-processing tasks [18]. Shi, Y et al. [19] have successfully integrated the U-KAN and C-KAN frameworks to address the challenge of medical image segmentation, accurately delineating anatomical structures and lesion areas, and providing a core basis for clinical decision-making. Omid Nejati Manzari et al. [20] embedded the KAN layer into Transformer to form a new framework and then proposed an enhanced Dilated Neighborhood Attention (DiNA) to achieve generalized medical image classification. Jianyun Liu et al. [21] devised a multi-scale feature-prediction network that first extracts feature maps via convolution, refines them with spatial attention, and finally sums the channels in a learnable Kolmogorov-Arnold layer to perform image classification.

B. Problem Statement

In the area of image processing, deep learning possesses substantial benefits, but the mainstream methods still have certain problems, mainly manifested as follows.

1) For CNN models, gradient degradation is prone to occur when the network deepens, making it difficult to fully capture global contextual information, and many model structures are complex.

2) CNNs can only represent linear mappings and are insensitive to higher-order statistics and texture information. In mammograms, however, malignant signs often manifest as higher-order distributions of texture and micro-calcification clusters. Relying solely on linear mappings makes it difficult to encode such high-order information into feature maps; additional nonlinear activations, pooling, or subsequent fully connected layers are required to partially compensate.

3) KAN adopts a fully connected topology, which has insufficient utilization of local priors in two-dimensional images and lacks spatial induction bias.

C. Objectives

To tackle these problems, this study proposes a KAN architecture enhanced with the Convolutional Block Attention Module (CBAM) [22] that successfully performs a three-class classification: benign, malignant, and normal. The main objectives are:

1) Introducing lightweight model(s) specifically designed for the classification of mammographic images is proposed.

2) Replacing the linear convolution kernels of traditional CNNs with KAN convolution kernels upgrades "fixed-weight linear filtering" to "nonlinear weight generation driven by learnable activation functions". This allows the model to conduct higher-order mapping modeling within convolution layers without increasing or even reducing the parameter scale, eliminating the parameter and computation explosion caused by subsequent stacking of multi-layer nonlinear activations and redundant channels.

3) Incorporating the CBAM technique into the convolutional layers based on KAN to build a new architecture, which is capable of adapting prominent regions in the images, efficiently reducing noise, extracting key local features, having faster convergence, and enhancing stability and accuracy.

This study integrates the advantages of the CBAM technique into the advantages of KAN-based models to build fusion models that inherit high feature extraction and strong interpretability, presenting successive architectures in the field of classification on mammography images. The new models are capable of adapting to images having different qualities, remarkably when facing images that are noisy and have low resolution, it still presents high accuracy and stability.

II. Related Studies

The proposed method embeds KAN's learnable univariate activation functions into the convolutional layer, instantiating nonlinear weights within the channel dimension to overcome the spatial insensitivity and limited expressiveness of conventional "linear kernels". CBAM further jointly produces attention coefficients across both channel and spatial axes, precisely highlighting diminutive lesions while suppressing background redundancy, thereby markedly elevating diagnostic accuracy and interpretability for mammographic images.

A. KAN

Kolmogorov-Arnold Network (KAN) equips each edge with a learnable activation function in place of conventional weights. Rather than fixed linear weights, KAN replaces every parameter with a univariate spline function. According to the KAN representation theorem, every continuous multivariate mapping can be decomposed into a collection of one-dimensional functions, as shown in Eq. (1):

$$f(X) = f(x_1, \ldots\ldots, x_n) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \emptyset_{q,p}(x_p) \right) \quad (1)$$

In a sense, they proved that the only true multivariate function is addition, because all other functions can be expressed through unary functions and summation operations. In a KAN, for a supervised-learning task made up of input-output pairs $\{x_i, y_i\}$, finding a function f so that y_i is approximately equal to $f(x_i)$ for each data point. The crux is to discover the appropriate univariate functions, inner (encoders) and outer (decoders); the architecture is shown in Fig. 1 (in the Appendix) [23].

Every learnable nonlinear function is univariate, taking only one input. It is modeled as a B-spline curve ϕ(x) whose local-basis coefficients c_i are learnable. In KAN, unlike traditional MLP, the weight of each connection in KAN is no longer a simple numerical value but is parameterized as a learnable spline function. By representing ∅_(q,p) and ϕ_q through B-splines and composing these functions, the entire network is constructed.

Overall, KAN, as a novel neural-network architecture, has demonstrated remarkable advantages both theoretically and experimentally, especially in accuracy, parameter efficiency, and interpretability.

B. CBAM

The Convolutional Block Attention Module (CBAM) is a simple and effective feedforward convolutional neural network attention module, which is commonly used for embedded int CNN to enhance the feature extraction capabilities from images. Given an intermediate feature map, the module infers the attention of the image along two independent dimensions, channel and space, and then multiplies the attention map by the input feature map for adaptive feature refinement [24][25][26][27][28]. CBAM is a lightweight universal module that can be well integrated into the KAN architecture presented in this article, with negligible overhead. This attention mechanism amplifies salient features while suppressing irrelevant ones.

Fig. 2 (in the Appendix) presents the architecture of the CBAM technique. As seen in this architecture, CBAM comprises two independent segments: CAM (Channel Attention Module) and SAM (Spatial Attention Module). Given an input feature map $F \in R^{(C\times H\times W)}$, CBAM sequentially produces a 1-D channel-attention $M_c \in R^{(C\times 1\times 1)}$, and a 2-D spatial-attention map $M_s \in R^{(1\times H\times W)}$ [29]. The overall attention procedure is summarized in Eq. (2) and (3) [30]. The ultimate output of CBAM is acquired by multiplying the result of the channel attention output by the weight of spatial attention.

$$F' = M_c(F) \otimes F \tag{2}$$

$$F'' = M_s(F') \otimes F' \tag{3}$$

1) CAM: Chanel Attention Module (CAM) at the first step squeezes the spatial axes of the feature map and then models cross-channel dependencies to deliver a channel-attention map. For the aggregation of spatial information, it adopts two types of pooling: max-pooling and average-pooling, which greatly improve the network's expressive power. CAM first gathers the three-dimensional information of the feature map via average-pooling and max-pooling, producing two distinct spatial-context descriptors, F^c_{avg} and F^c_{max}. These summaries are then fed to a single Multilayer Perceptron (MLP) subnet that outputs the channel-attention map $M_c \in \{\mathbb{R}^{c\times 1\times 1}\}$. The equation for calculating CAM is presented in Eq. (4) [26]. σ represents the Sigmoid function, W_0 represents the dimensionality reduction weight matrix with a shape of (c/r)×c, and W_1 represents the dimensionality enhancement weight matrix with a shape of c×(c/r), where r is considered the rate of reduction and the default value is 16.

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{4}$$

2) SAM: Spatial Attention Module (SAM) multiplies CAM's channel attention by the input feature map to get SAM's input feature map. Spatial attention is different from channel attention, as it focuses on the type of information within a Feature Map and complements channel attention. Use two *channel=c* pooling operations to aggregate the channel information of feature maps, generating two 2-D feature maps. Each feature map aggregates channel information, representing the average and max-pooling characteristics across channels. Then use a standard convolutional layer to connect and convolve them to generate a 2-D spatial attention map, with the calculation shown in Eq. (5).

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{5}$$

Finally, a 7×7 convolution is applied to produce the spatial-attention matrix. Then, the sigmoid function converts the response map into spatial-attention weights, called the spatial attention map $M_s(F)$. Finally, the input feature map with the channel attention mechanism is element-wise multiplied by the output $M_s(F)$ to obtain the result of CBAM.

III. Methodology

This study proposes a lightweight model for mammographic image classification based on KAN-CBAM. By leveraging KAN, the network overcomes the nonlinear-classification bottlenecks of image processing while remaining highly interpretable. To further enhance feature extraction, we embed the CBAM attention mechanism inside every KAN convolutional layer, which not only helps the model converge faster but also helps extract more useful image features, focus on discriminative features, and make classification more accurate. Fig. 3 (in the Appendix) illustrates the overall architecture of KAN-CBAM.

The proposed framework is a convolutional neural architecture that combines the KAN and CBAM mechanisms. It consists of two convolutional blocks (Conv Block) and one classifier. Each Conv Block performs two key operations, namely KAN convolution and CBAM. The result from Conv Block 1 is channeled into Conv Block 2 through a residual connection, which helps address the vanishing gradient problem in deep networks and enhances the integrity of feature propagation. Finally, the Classifier module converts the extracted features into class probabilities via fully connected layers, batch normalization, and Dropout, completing the classification task.

Due to the introduction of KAN-enhanced feature extraction, combined with CBAM-enhanced feature selection and the use of residual learning to alleviate gradient degradation, the proposed model improved representation ability while reducing redundancy. Finally, it was validated on the dataset, achieving efficient and high-precision image classification.

A. KAN Convolution

According to Fig. 3, the KAN-Convolutional Layer is located in Conv Blocks. The KAN convolutional layer extracts features with its unique function approximation ability, providing more expressive initial feature maps for subsequent processing.

KAN convolution is a special convolution operation that applies a learnable nonlinear function to each edge and adds them together. Every weight is recast as a univariate spline function, and nodes merely add up the input signals without using nonlinear operations.

Traditional CNNs mainly depend on set activation functions and linear transformations to capture spatial relationships. Although this design effectively handles complex spatial data, it lacks sufficient flexibility and expressiveness. By directly replacing the nonlinear activation function of KANs in the convolutional layer, KAN convolutional layers can be constructed. KANs reduce the number of parameters and may improve generalization ability by introducing learnable spline functions, B-splines. This new layer type not only maintains the efficiency of CNN in processing image data but also enhances its flexibility and expressiveness through the spline function of KANs.

In KAN convolution, the spline is constructed primarily from B-spline basis functions, as given in Eq. (6), where $B_i(x)$ represents a B-spline basis function, and c_i is a learnable coefficient that controls the weights of the basis functions. The spline is constructed by linearly combining multiple local basis functions, and the parameters that KAN truly needs to learn are the coefficients c_i in front of the basis functions. By adjusting the parameters of the spline, it can adapt to different data distributions and task requirements.

$$spline(x) = \sum_{i=1}^{n} c_i \cdot B_i \quad (6)$$

B. KAN-CBAM Model

To enhance the automatic focusing ability of KAN convolutional layers on discriminative regions and channels, the model inserts the Convolutional Block Attention Module (CBAM) after KAN convolution to form KANConv-CBAM units. CBAM applies channel and spatial dimension attention weighting to the feature maps of channels 4 and 8, respectively, highlighting key features, suppressing irrelevant information, and enhancing the effectiveness of feature expression.

The channel attention structure is shown in Fig. 4 (in the Appendix), which chiefly employs channel management among features to create channel attention maps. In the module, the input feature map F, which has dimensions of C×H×W, experiences both global max-pooling and global average-pooling operations. Specifically, global max-pooling is carried out along the W dimension, while global average-pooling is performed along the H dimension. Consequently, two feature maps with dimensions of C×1×1 are acquired. These maps are subsequently input into a 2-layer MLP, in which the output features are aggregated and sigmoid-activated to produce the final feature.

The architecture of the spatial attention module is shown in Fig. 5 (in Appendix), which generates a spatial attention map using the spatial relationships between features, focusing on where the information is located. This module utilizes the feature map, which is the output of the Channel Attention Module, as its input. The feature map undergoes average-pooling and max-pooling operations to obtain two feature maps. Next, the two feature maps are merged in the channel dimension and connected and convolved using a standard convolution layer to obtain a 2-dimensional spatial attention map. Finally, the sigmoid function is used to achieve spatial attention features.

This spatial attention map reflects the importance of different spatial positions in the feature map. It is then element-wise multiplied by the channel-attended feature map, spatially re-weighting the activations so that salient locations are emphasized.

IV. Results and Discussion

This section first introduces the datasets and then presents the experiment parameters. Ultimately, the results of the experiment are compared and analyzed to confirm the effectiveness of the algorithm proposed in this study.

A. Datasets

This study mainly selects to validate the model using the public datasets: DDSM, which includes 5,238 images, INbreast, which includes 13,172 images, and MIAS, which includes 322 images. These datasets are mainly divided into three categories: normal, benign, and malignant. Each dataset was partitioned with an 80/20 split for training and testing.

Through analysis of the datasets, it is known that there are many images with inconsistent resolutions, noise interference, and significant differences in image size and shape, as shown in Fig. 6 (a),(b), and(c) (in Appendix). However, the proposed method effectively overcomes these problems in the dataset and efficiently completes the image classification task.

B. Experimental Environment

To verify the effectiveness of this project, validation was conducted on a laptop computer and in the hardware environment as listed in Table I.

TABLE I. Hardware Environment Parameters

Hardware	Parameters
GPU	NVIDIA T600
RAM	32.0 GB
Storage	2TB
CPU	Intel Core i7-11700 @ 2.50GHz

C. Experimental Results

1) Major metrics comparison: This study primarily employs accuracy, loss, F1-score, Recall, and AUC as major evaluation metrics to verify the performance of the model. It mainly compares the classification performance of seven KAN-based models and KAN-CBAM models enhanced with the CBAM. Table VI (in the Appendix) presents the

classification results of the seven KAN-based models, while Table VII (in the Appendix) shows those of the seven KAN-CBAM models. In the tables, "M" stands for Medium, "B" stands for Big, and "S" stands for Small.

Experimental results in Table VI and Table VII (in the Appendix) indicate KAN-CBAM-based models exhibit superior performance – especially regarding accuracy, F1 score and AUC – compared to KAN-based models among the three datasets.

On the DDSM dataset, in KAN-CBAM-based models, KANC MLP (B) is the superior model, presenting values 84.85%, 84.01%, and 85.87% regarding accuracy, F1-score, and AUC, respectively, while in KAN-based models, KKAN (M)is the superior model, showing values 84.04%, 83.09%, and 84.15%.

On the INbreast dataset, in KAN-CBAM-based models, KKAN (M) is the superior model, presenting values of 99.66%, 99.64%, and 99.70% regarding accuracy, F1-score, and AUC, respectively. In KAN-based models, KANC MLP (B) is the superior model for accuracy and F1-score, showing values 98.90% and 98.82%, respectively, while KANC MLP (M) is the superior model for AUC, having a value of 99.62%.

On the MIAS dataset, in KAN-CBAM-based models, KKAN (S) is the superior model regarding accuracy and AUC, having values of 85.00% and 79.21%, respectively, and KANC MLP (M) is the superior model regarding F1-score, having a value of 83.33%. In KAN-based models, KKAN (S) is the superior regarding accuracy, F1-score, and AUC, having values of 75.00%, 49.36%, and 59.21%, respectively.

2) AUC comparison: To evaluate the effect of the CBAM technique on the KAN-based models, AUC values of all KAN-based models were compared with corresponding KAN-CBAM-based models for DDSM, INbreast, and MIAS datasets separately. Fig. 7 (in Appendix), Fig. 8 (in Appendix), and Fig. 9 (in Appendix) present AUC comparison on DDSM, INbreast, and MIAS datasets, respectively.

As Fig. 7 shows, for the DDSM dataset, generally, KAN-CBAM-based models accomplished higher AUC values compared to KAN-based models. The most enhancement was for the KANC (M) model, in which the AUC value increased from 51.27% to 81.13%, introducing a strong improvement in discriminative potential after integrating the CBAM technique. Further enhancements were also seen for KAN (M), KAN (S), KANC (B), KANC (S), and KKAN (S). Only KKAN (M) presented a trivial reduction in AUC.

For INbreast dataset, KAN-based models and KAN-CBAM-based models both made high AUC values, signaling that this dataset is extremely separable for most models. Therefore, the differences in AUC values between KAN-based models and KAN-CBAM-based models are quite small.

For the MIAS dataset, the contribution of the CBAM technique was the brightest, since all KAN-CBAM-based models overtook their relevant KAN-based models. The largest increases in AUC were seen in KANC (B), KANC (S), KKAN (M), and KKAN(S), indicating that the CBAM technique affected highly feature enhancement and class discrimination on this dataset.

Taken together, AUC analysis shows that generally the CBAM technique enhances discriminative performance, with the highest effect seen on the MIAS dataset, followed by the DDSM dataset, while on INbreast dataset, the contribution of CBAM is trivial due to the existence of high reference performance.

3) Loss comparison: A comparison was performed over loss values of KAN-based models and KAN-CBAM-based models, since smaller loss indicates better optimization behavior. Fig. 10 (in Appendix), Fig. 11 (in Appendix), and Fig. 12 (in Appendix) present the loss comparison of DDSM, INbreast, and MIAS datasets, respectively.

For the DDSM dataset, generally, KAN-CBAM-based models present lower loss values compared to relevant KAN-based models. The largest reduction was seen for KANC (M), in which the loss value dropped from a value of 0.0291 to a value of 0.0115. Another large reduction was seen for KKAN (M), in which the loss decreased from a value of 0.0118 to a value of 0.0030.

For INbreast dataset, KAN (M), KAN (S), and KANK (S) introduced slightly lower loss values without the CBAM technique, while KANC (M) showed a large improvement after the incorporation of the CBAM technique. For the remaining models, the differences were trivial or zero.

For the MIAS dataset, the effect of the CBAM technique was not consistent. A small reduction was seen for KAN (M), whereas KKAN (M and KKAN (S) persisted unchanged. All other models presented trivial growth in loss value after the integration of the CBAM technique. This specifies that the effect of CBAM on loss optimization is architecture-dependent.

Overall, loss analysis shows that the CBAM technique presents the strongest optimization on the DDSM dataset, while its effect on INbreast and MIAS datasets is not consistent and depends on model configuration.

4) Other metrics comparison: Precision, specificity, and Matthew Correlation Coefficient (MCC) capture complementary aspects and attributes of analytical reliability which are not exhibited by accuracy, F1-score, recall, AUC, and loss.

Precision is important since it shows the ratio of predicted positive samples that are truly identified as positive. In this study, greater precision shows smaller false positive predictions, thus helping decrease unnecessary follow-up tests and patient fear and anxiety.

Specificity shows the ratio of predicted negative samples that are truly negative, indicating the model's capability to prevent falsely classifying normal cases or benign cases as abnormal.

MCC presents a well-adjusted summary of the performance of the classification by considering four parameters in the confusion matrix: true positive, false positive, true negative,

and false negative. The summary presented by MCC is helpful when a complete view of the reliability of a model in classification is needed.

To prevent an extremely long presentation, precision, specificity, and MCC are summarized as average values shown in Table II.

As Table II shows, generally, KAN-CBAM-based models attain larger average values than KAN-based models within all three metrics (precision, specificity, and MCC) across all datasets. This supports the idea that the integration of the CBAM technique not only enhances discrimination performance but also increases the reliability of the classification.

5) Delta analysis: To have a better numerical understanding of the effect of the CBAM technique, the absolute changes between each KAN-based model and the corresponding KAN-CBAM-based model were computed for AUC and loss. It should be noted that a positive value for ΔAUC means that the CBAM technique enhanced discriminative performance, while a negative value for ΔLoss means that the CBAM technique lowered the loss value, hence enhancing optimization behavior. Tables III, IV, and V present absolute changes after the inclusion of the CBAM technique on DDSM, INbreast, and MIAS datasets, respectively.

The values in the delta analysis support the values in the bar charts presented earlier. On the DDSM dataset, the CBAM technique enhances most models, while the largest benefit is seen for KANC (M). On INbreast dataset, the contribution of CBAM was minor and somewhat mixed, which is likely associated with the high existing performance of KAN-based models. On the MIAS dataset, all models gained from the CBAM technique in terms of AUC, implying a consistent positive effect of architecture-based feature enhancement on this dataset.

The values related to ΔLoss also prove that enhancements in AUC are not always consistent with reductions in loss. For instance, CBAM showed a large loss reduction on the DDSM dataset, whereas its effect on MIAS and INbreast datasets was dependent on the model's architecture. This proposes that the contribution of the CBAM technique is not solely optimization-driven but also relevant to how successfully the attention process improves discriminative feature interpretation.

TABLE II. SUMMARY OF PRECISION, SPECIFICITY, AND MCC

Dataset (Model based)	Avg. Precision	Avg. Specificity	Avg. MCC
DDSM (KAN)	0.734	0.747	0.484
DDSM (KAN-CBAM)	0.800	0.813	0.604
INbreast (KAN)	0.876	0.969	0.864
INbreast (KAN-CBAM)	0.974	0.989	0.970
MIAS (KAN)	0.414	0.553	0.141
MIAS (KAN-CBAM)	0.813	0.831	0.616

TABLE III. ABSOLUTE CHANGE AFTER CBAM INTEGRATION ON DDSM DATASET

Model Pair	ΔAUC	ΔLoss
Conv & KAN (M)	+0.0712	-0.0009
Conv & KAN (S)	+0.0308	-0.0005
KANC MLP (B)	+0.0358	-0.0006
KANC MLP (M)	+0.2986	-0.0176
KANC MLP (S)	+0.0466	-0.0013
KKAN (M)	-0.0036	-0.0088
KKAN (S)	+0.0134	-0.0004

TABLE IV. ABSOLUTE CHANGE AFTER CBAM INTEGRATION ON INBREAST DATASET

Model Pair	ΔAUC	ΔLoss
Conv & KAN (M)	-0.0173	+0.0124
Conv & KAN (S)	-0.0305	+0.0167
KANC MLP (B)	+0.0002	+0.0001
KANC MLP (M)	-0.0017	-0.0004
KANC MLP (S)	+0.0236	+0.0005
KKAN (M)	+0.0314	+0.0009
KKAN (S)	+0.0219	+0.0015

TABLE V. ABSOLUTE CHANGE AFTER CBAM INTEGRATION ON MIAS DATASET

Model Pair	ΔAUC	ΔLoss
Conv & KAN (M)	+0.2093	-0.0037
Conv & KAN (S)	+0.1916	+0.0040
KANC MLP (B)	+0.2836	+0.0033
KANC MLP (M)	+0.1882	+0.0005
KANC MLP (S)	+0.2296	+0.0042
KKAN (M)	+0.2000	+0.0000
KKAN (S)	+0.2000	+0.0000

D. Summary of Performance Indicators

The results show that the inclusion of the CBAM technique generally enhances KAN-based models' performance within the datasets considered. This matter is strongly confirmed not only by AUC, loss, and delta analyses, but also by other metrics such as precision, specificity, and MCC. Overall, the findings denote that the CBAM technique offers not only higher discriminative capability but also greater enhancement in classification reliability.

V. CONCLUSION AND FUTURE WORKS

This study worked on the effect of the CBAM technique by comparing KAN-based and KAN-CBAM-based models on the three datasets – DDMS, INbreast, and MIAS – including breast cancer images.

The results indicated that incorporating the CBAM technique enhances performance, especially regarding AUC, in which the most consistent benefits were seen on the MIAS

dataset, and high enhancement was found on the DDSM dataset, whereas the improvement was trivial on INbreast dataset, since most of the models already attained very high performance.

Based on the results of the loss analysis, the contribution of the CBAM technique to optimization likely depends on the architecture and the dataset. The highest loss reduction was seen on the DDSM dataset, whereas the impacts on the MIAS and INbreast datasets were mixed. Moreover, the results regarding precision, specificity, and MCC confirmed the advantage of integrating CBAM into the models studied.

In future studies, the models studied can be assessed under other testing settings and on other medical datasets. In addition, future studies can integrate robustness analysis and statistical significance testing over explanatory models to have a better evaluation of CBAM and its effects.

REFERENCES

[1] https://who.int/news-room/fact-sheets/detail/breast-cancer, Accessed on: January 6, 2026.

[2] http://www.52qixiang.com/info/83134.html, Accessed on: January 6, 2026.

[3] S. K. Biswas and D. P. Mukherjee, "Recognizing architectural distortion in mammogram: A multiscale texture modeling approach with GMM," IEEE Trans. Biomed. Eng., vol. 58, no. 7, pp. 2023-2030, July 2011, DOI: 10.1109/TBME.2011.2128870

[4] L. Balkenende, J. Teuwen, and R. M. Mann, "Application of deep learning in breast cancer imaging," Seminars in Nuclear Medicine, vol. 52, no. 5, pp. 584-596. March 2022, DOI: 10.1053/j.semnuclmed.2022.02.003

[5] S. H. Heywang-Köbrunner, A. Hacker, and S. Sedlacek, "Advantages and disadvantages of mammography screening," Breast Care, vol. 6, no. 3, pp. 199–207, May 2011, DOI: 10.1159/000329005

[6] Z. Gandomkar and C. Mello-Thoms, "Visual search in breast imaging," Br. J. Radiol., vol. 92, no. 1102: 20190057, July 2019, DOI: 10.1259/bjr.20190057

[7] L. A. Aldakhil, H. F. Alhasson, S. S. Alharbi, R. U. Khan and A. M. Qamar, "Image-based breast cancer histopathology classification and diagnosis using deep learning approaches," Appl. Comput. Intell. Soft Comput., vol. 2025, Art. no. 7011984, February 2025, DOI: 10.1155/acis/7011984

[8] S. Islam et al., "Image processing and Support Vector Machine (SVM) for classifying environmental stress symptoms of pepper seedlings grown in a plant factory," Agronomy, vol. 14, no. 9, Art. no. 2043, September 2024, DOI: 10.3390/agronomy14092043

[9] M. Hosseinpour, S. Ghaemi, S. Khanmohammadi and S. Daneshvar, "A hybrid high-order type-2 FCM improved random forest classification method for breast cancer risk assessment," Appl. Math. Comput., vol. 424, p. 127038, July 2022, DOI: 0.1016/j.amc.2022.127038

[10] A. Igali and P. Shamoi, "Image classification using fuzzy pooling in convolutional Kolmogorov-Arnold networks," in SCIS & ISIS 2024, Himeji, Japan, 2024, pp. 1-7

[11] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 53, March 2021, DOI: 10.1186/s40537-021-00444-8

[12] S. Alam Ansari, A. P. Agrawal, M. A. Wajid, M. S. Wajid and A. Zafar, "MetaV: A pioneer in feature augmented meta-learning based vision transformer for medical image classification," Interdiscip. Sci. Comput. Life Sci., vol. 16, No. 2, pp. 469-488, June 2024, DOI: 10.1007/s12539-024-00630-1

[13] Z. Jafari and E. Karami, "Breast cancer detection in mammography images: A CNN-based approach with feature selection," Information, vol. 14, no. 7, 410, July 2023, DOI: doi.org/10.3390/info14070410

[14] L. Bouzar-Benlabiod, K. Harrar, L. Yamoun, M. Y. Khodja and M. A. Akhloufi "A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification," Comput. Biol. Med., vol. 163, p. 107133, June 2023, DOI: 10.1016/j.compbiomed.2023.107133

[15] S. Majumdar, P. Pramanik and R. Sarkar, "Gamma function based ensemble of CNN models for breast cancer detection in histopathology images," Expert Syst. Appl., vol. 213, Part B, 119022, March 2023, DOI: 10.1016/j.eswa.2022.119022

[16] J. G. Melekoodappattu, A. S. Dhas, B. K. Kandathil and K. S. Adarsh, "Breast cancer detection in mammogram: Combining modified CNN and texture feature based approach," J. Ambient Intell. Human. Comput., vol. 14, pp. 11397-11406, September 2023, DOI: 10.1007/s12652-022-03713-3

[17] J. Gao et al., "MSFKAN: A multi-scale feature prediction network combined with KAN for medical image classification," Neural Process. Lett., vol. 57, no. 67, July 2025, DOI: 10.1007/s11063-025-11782-6

[18] N. Firsov et al., "HyperKAN: Kolmogorov–Arnold networks make hyperspectral image classifiers smarter," Sensors, vol. 24, no. 23, 7683, November 2024, DOI:10.3390/s24237683

[19] Y. Shi, S. Jiang, X. Liu, Y. Gao and G. Xu, "U-CKAM based on C-KAN and MRF for medical image segmentation," in Data Inf. Online Environ. (DIONE 2024), Lect. Notes Inst. Comput. Sci., Soc. Inform. Telecommun. Eng., vol. 569, Springer, Cham, 2026, pp. 235-246, DOI: 10.1007/978-3-031-97352-9_19

[20] O. Nejati Manzari, H. Asgariandehkordi, T. Koleilat, Y. Xiao and H. Rivaz, "Medical image classification with KAN-integrated transformers and dilated neighborhood attention," Appl. Soft Comput., vol. 178, 114045, October 2025, DOI: 10.1016/j.asoc.2025.114045

[21] T. Liu, Y. Yu, M. Yang, Z. Xiao, Q. Li and Y. Tang, "LCTKAN: Lightweight convolution Transformer–KAN for medical image classification," Biomedical Signal Processing and Control, vol. 116, 109553, May 2026, DOI: 10.1016/j.bspc.2026.109553

[22] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in ECCV, Lecture Notes in Computer Science, vol. 11211, Cham, Switzerland: Springer, 2018, pp. 3-19, DOI: 10.1007/978-3-030-01234-2_1

[23] Z. Liu et al., "KAN: Kolmogorov-Arnold networks," in Int. Conf. Learn. Representations (ICLR), vol. 2025, pp. 70367-70413

[24] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, September 2018, pp.3-19, DOI: 10.1007/978-3-030-01234-2_1

[25] M. Zakariah and A. Alnuaim, "Recognizing human activities with the use of Convolutional Block Attention Module," Egyptian Informatics Journal, vol. 27, September 2024, 100536, DOI: 10.1016/j.eij.2024.100536

[26] J. Qin, S. Zhang, Y. Wang, F. Yang, X. Zhong and W. Lu, "Improved skeleton-based activity recognition using convolutional block attention module," Computers and Electrical Engineering, vol. 116, May 2024, 109231, DOI: 10.1016/j.compeleceng.2024.109231

[27] M. M. Farag, M. Fouad and A. T. Abdel-Hamid, "Automatic Severity Classification of Diabetic Retinopathy Based on DenseNet and Convolutional Block Attention Module," IEEE Access, vol. 10, pp. 38299-38308, 2022, DOI: 10.1109/ACCESS.2022.3165193

[28] S. Agac and O. D. Incel, "On the Use of a Convolutional Block Attention Module in Deep Learning-Based Human Activity Recognition with Motion Sensors," Diagnostics, vol. 13, no. 11, 1861, May 2023, DOI: 10.3390/diagnostics13111861

[29] P. Wen, et al. "A-PSPNet: A novel segmentation method of renal ultrasound image", IEEE International Conference on Systems, Man, and Cybernetics (SMC), October 2021, pp. 40-45, DOI: 10.1109/SMC52423.2021.9658740

[30] K. Zheng, D. Yang, J. Liu and J. Cui. "Recognition of teachers' facial expression intensity based on convolutional neural network and attention mechanism", IEEE Access, vol. 8, pp. 226437-226444, December 2020, DOI: 10.1109/ACCESS.2020.3046225

APPENDIX

TABLE VI. EXPERIMENTAL RESULTS OF KAN-BASED MODELS

Dataset	Evaluation metric	Conv & KAN (M)	Conv & KAN (S)	KANC MLP (B)	KANC MLP (M)	KANC MLP (S)	KKAN (M)	KKAN (S)
DDSM	Accuracy	0.7459	0.7125	0.8225	0.5132	0.8029	0.8404	0.8111
	Loss	0.0141	0.0152	0.0121	0.0291	0.0126	0.0118	0.0122
	F1-score	0.7075	0.6958	0.8193	0.5012	0.8016	0.8309	0.7996
	Recall	0.7226	0.7051	0.8202	0.5996	0.7916	0.8371	0.8103
	AUC	0.7011	0.6928	0.8229	0.5127	0.8023	0.8415	0.8105
INbreast	Accuracy	0.9620	0.9620	0.9890	0.9838	0.8368	0.8102	0.7443
	Loss	0.0001	0.0001	0.0005	0.0010	0.0001	0.0001	0.0001
	F1-score	0.9589	0.9589	0.9882	0.9825	0.8033	0.7649	0.6553
	Recall	0.9571	0.9571	0.9873	0.9796	0.7824	0.7460	0.6550
	AUC	0.9862	0.9862	0.9956	0.9962	0.9715	0.9656	0.9748
MIAS	Accuracy	0.7153	0.7307	0.7307	0.7307	0.7307	0.7307	0.7500
	Loss	0.0512	0.0435	0.0398	0.0387	0.0416	0.0378	0.0865
	F1-score	0.4089	0.4222	0.4222	0.4222	0.4222	0.4222	0.4936
	Recall	0.4762	0.5000	0.5000	0.5000	0.5000	0.5000	0.5357
	AUC	0.5218	0.5395	0.4586	0.5733	0.5000	0.5508	0.5921

TABLE VII. EXPERIMENTAL RESULTS OF KAN-CBAM-BASED MODELS

Dataset	Evaluation metrics	Conv & KAN (M)	Conv & KAN (S)	KANC MLP (B)	KANC MLP (M)	KANC MLP (S)	KKAN (M)	KKAN (S)
DDSM	Accuracy	0.8046	0.7769	0.8485	0.8339	0.8404	0.8371	0.8257
	Loss	0.0132	0.0147	0.0115	0.0115	0.0113	0.0030	0.0118
	F1-score	0.8013	0.7727	0.8401	0.8263	0.8339	0.8258	0.8241
	Recall	0.7916	0.7572	0.8503	0.8305	0.8471	0.8376	0.8227
	AUC	0.7723	0.7236	0.8587	0.8113	0.8489	0.8379	0.8239
INbreast	Accuracy	0.9821	0.9784	0.9886	0.9862	0.9914	0.9966	0.9924
	loss	0.0125	0.0168	0.0006	0.0006	0.0006	0.0010	0.0016
	F1-score	0.9359	0.9206	0.9877	0.9852	0.9908	0.9964	0.9918
	Recall	0.9219	0.9707	0.9883	0.9870	0.9916	0.9968	0.9913
	AUC	0.9689	0.9557	0.9958	0.9945	0.9951	0.9970	0.9967
MIAS	Accuracy	0.8076	0.8076	0.8230	0.8461	0.8153	0.8307	0.8500
	Loss	0.0475	0.0475	0.0431	0.0392	0.0458	0.0378	0.0865
	F1-score	0.7973	0.7973	0.8148	0.8333	0.8065	0.8222	0.7936
	Recall	0.7256	0.7256	0.7408	0.7642	0.7319	0.7521	0.7357
	AUC	0.7311	0.7311	0.7422	0.7615	0.7296	0.7508	0.7921

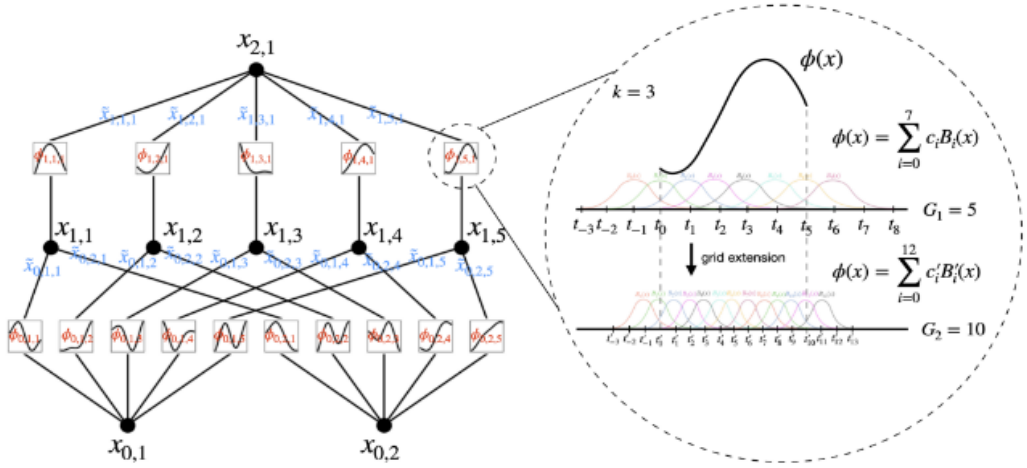


Fig. 1. KAN structure diagram.

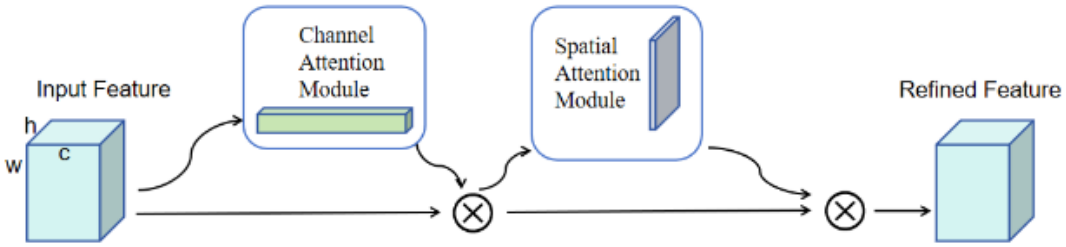


Fig. 2. CBAM architecture.

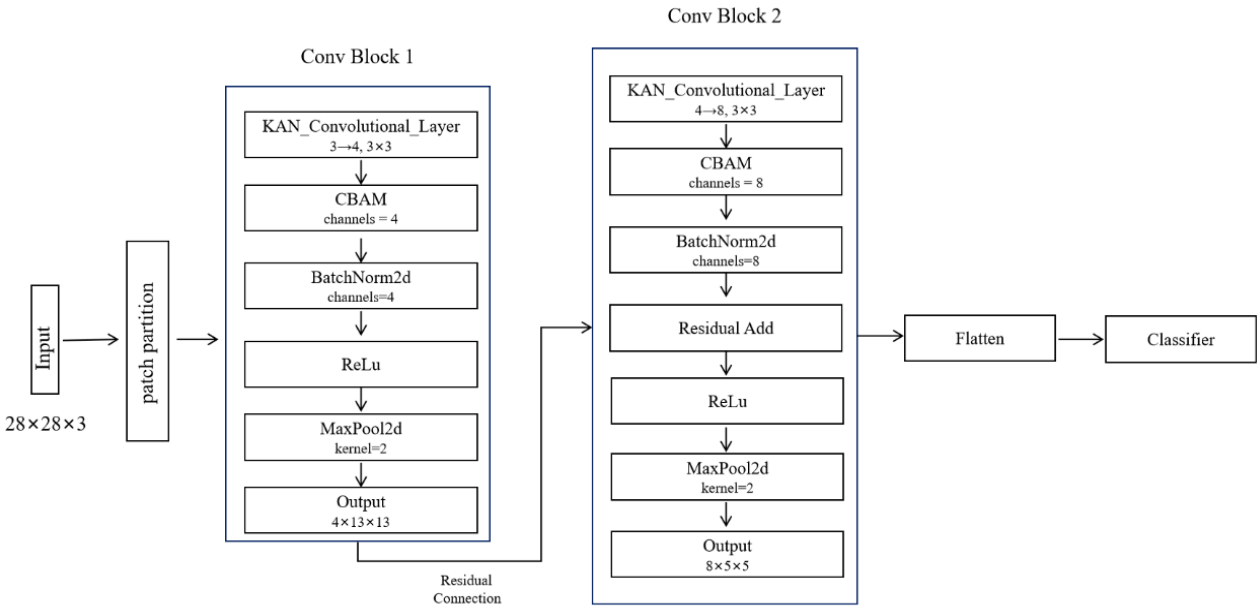


Fig. 3. Architecture of mammographic image classification model based on KAN-CBAM.

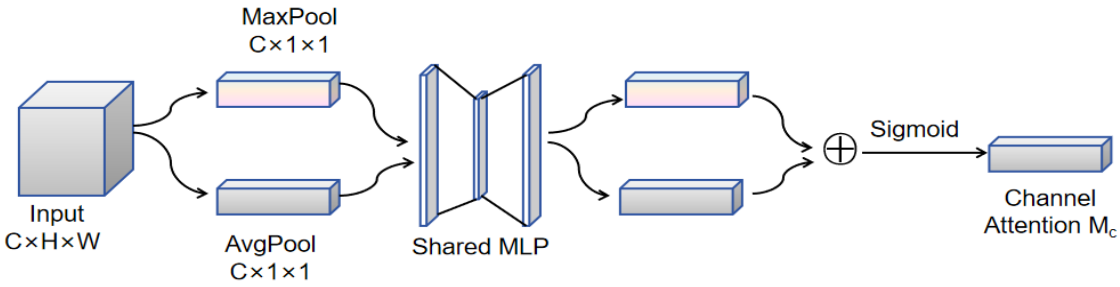


Fig. 4. CAM (Channel Attention Module).

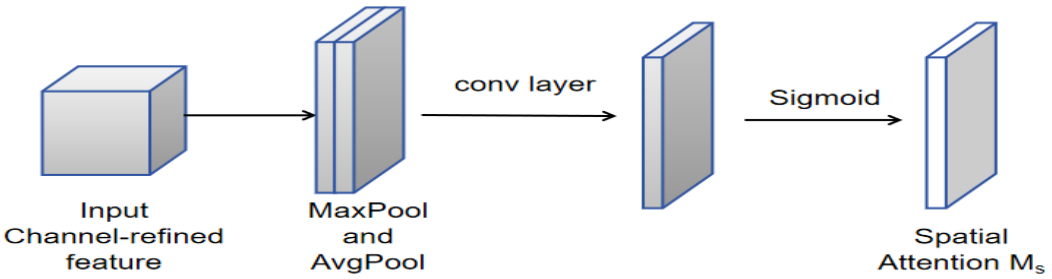
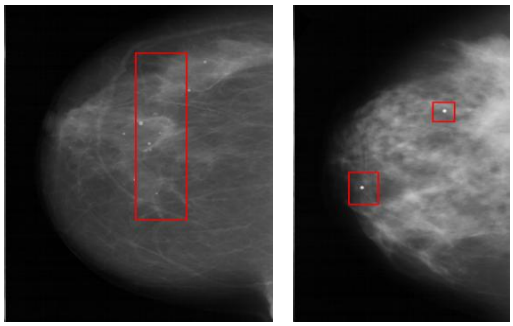
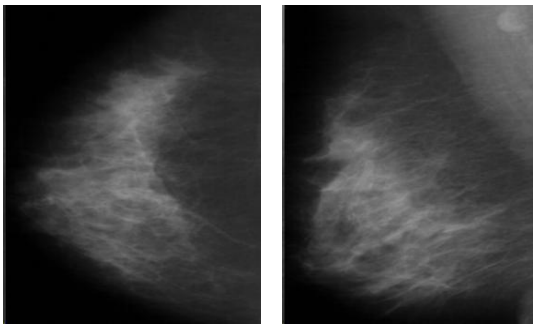


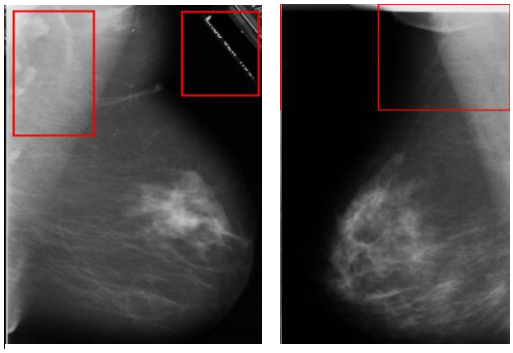
Fig. 5. SAM (Spatial Attention Module).



(a) Bright spots are easily misidentified as microcalcifications.



(b) The image resolution is relatively low.



(c) Image noise interference.

Fig. 6. Dataset samples.

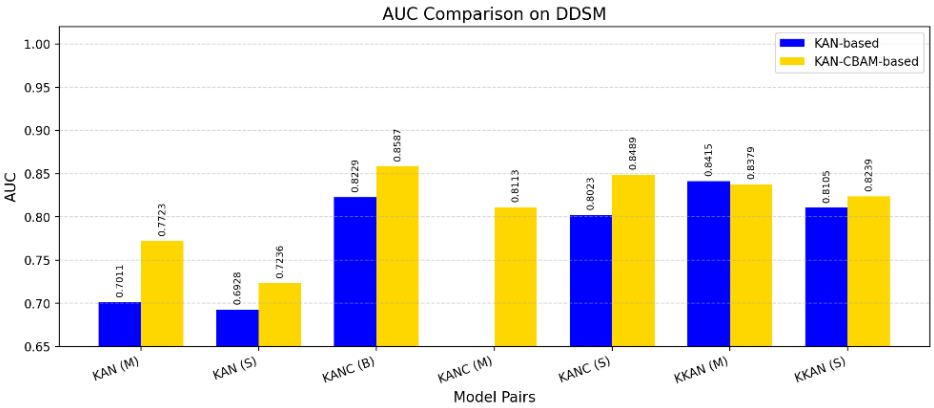


Fig. 7. AUC comparison of KAN-based and KAN-CBAM-based models for DDSM dataset.

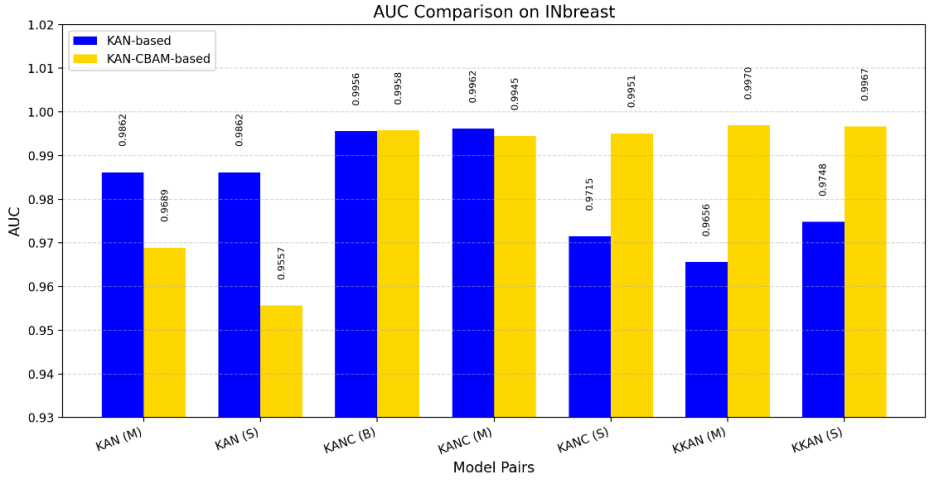


Fig. 8. AUC comparison of KAN-based and KAN-CBAM-based models for INbreast dataset.

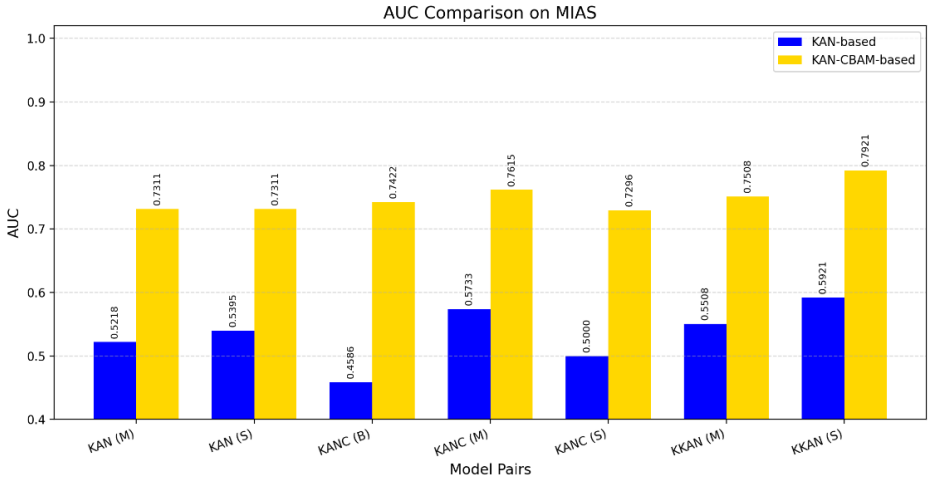


Fig. 9. AUC comparison of KAN-based and KAN-CBAM-based models for MIAS dataset.

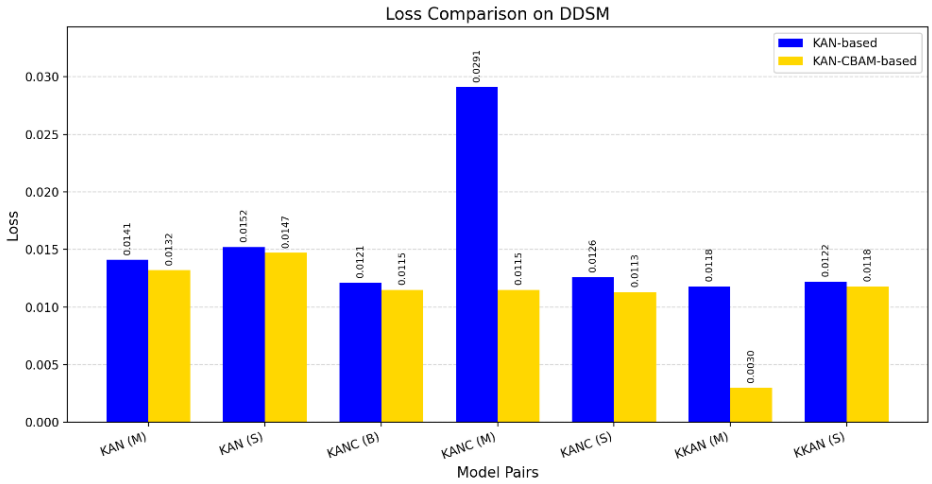


Fig. 10. Loss Comparison of KAN-based and KAN-CBAM-based models for DDSM dataset.

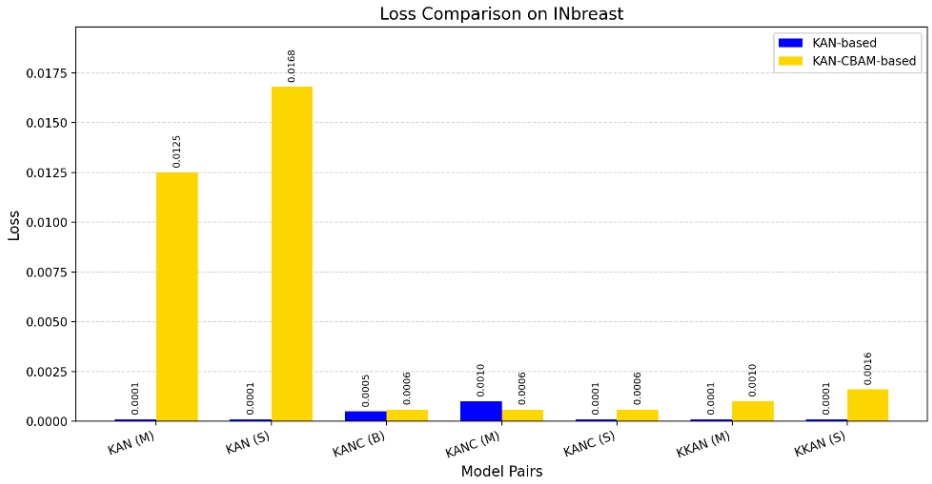


Fig. 11. Loss comparison of KAN-based and KAN-CBAM-based models for INbreast dataset.

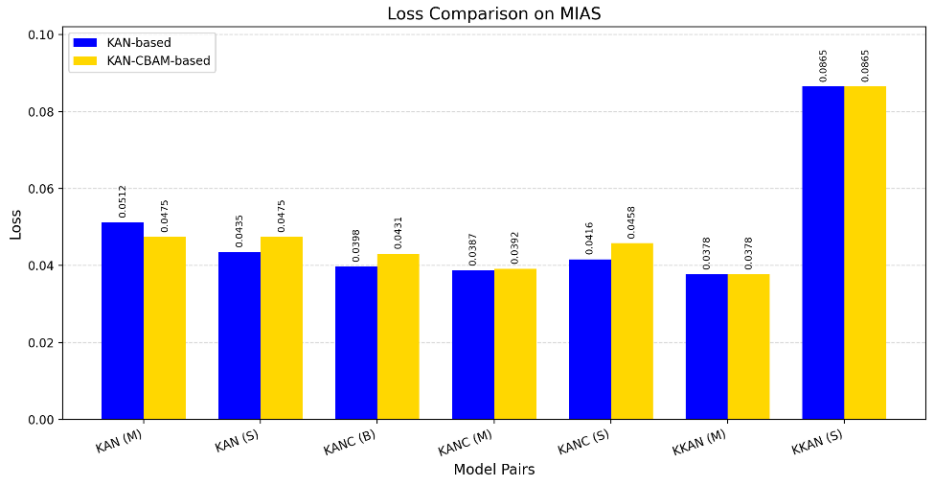


Fig. 12. Loss comparison of KAN-based and KAN-CBAM-based models for MIAS dataset.

Enhancing Virtual Team Performance Through a Quantitative-Based Management Model (VTPM)

Sathya A/L Nantha Kumar, Zulkefli Mansor
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*—The rapid growth of globalization, digital transformation, and remote working practices has significantly increased the reliance on virtual teams within modern organizations. Despite their advantages, virtual teams continue to face challenges such as communication barriers, inadequate technology management, unclear role allocation, ineffective strategic planning, and excessive documentation practices, all of which negatively affect team performance. This study aims to identify the key factors affecting virtual team performance and to develop a structured Virtual Team Performance Management (VTPM) model to improve the effectiveness of virtual collaboration. A quantitative research approach was adopted, involving questionnaire-based data collection with 110 respondents who had worked in virtual teams (instrument reliability testing was conducted using Cronbach's Alpha analysis). The findings indicate that role clarity, effective communication, effective strategic planning, task allocation, and appropriate management of technology are important in the determination of the performance of virtual teams. Specifically, 39.1% of respondents rated communication as "needs improvement," 63.6% agreed that excessive documentation reduces efficiency, and 56.4% strongly agreed that strategic planning is effective. On this premise, the Virtual Team Performance Management Model (VTPM) has been constituted and validated by three academic and industry experts through expert evaluation. The model received positive evaluations from academic and industry experts regarding its practicality and relevance in virtual team environments. This study provides a comprehensive and empirically informed model that organizations can apply to improve the work of virtual teams.**

***Keyword*—Virtual teams; team performance; remote work; VTPM model; communication; quantitative study**

I. Introduction

The latest dramatic shifts to digital technology and overwhelming effects of globalization have created and triggered the enormous application of virtual teams, groups of people that are physically separated and communicate using digital media. This change was also triggered by the onset of the COVID-19 pandemic [7], [32]. Among these advantages, virtual teams introduce the following, which include flexibility, access to global talent, and reduced cost of operation [32]. Nevertheless, they suffer significant performance challenges, which consist of communication barriers, leadership challenges, and the overuse of digital and asynchronous tools [5], [19], [32], the ambiguity of the roles that lead to the undermining of accountability [11], [43], which is determined by the absence of technology use or insufficient training [42]. Other than this, the absence of systematic strategic planning can lead to long delays in the processes and ineffective communication [22].

Scope boundary of the VTPM model: The VTPM model is empirically informed for technology-mediated, task-oriented virtual teams operating in corporate, public, and academic sectors. It assumes stable internet connectivity, basic digital literacy among team members, and the use of standard collaboration tools (e.g., Microsoft Teams, Zoom, Slack). The model may not be directly applicable to purely social virtual communities, emergency response teams, highly creative design teams where unstructured communication is critical, or teams in low-connectivity environments without adaptation. Fig. 1 illustrates the key challenges affecting virtual team performance identified from the literature review and empirical observations.



Fig. 1. Key challenges affecting virtual team performance.

Despite the numerous models of performance, a large part of performance models has not been empirically informed and has no implication to be applied practically to virtual teams in the actual real-life context [22], [27]. In one of the attempts to fill this knowledge gap, the suggested research offers the Virtual Team Performance Management Model (VTPM). The model is aimed at fulfilling three clear objectives, namely, to recognize the issues of virtual teams, to recognize the key factors that influence their performance, and to develop a management framework supported by empirical findings, based on the received results.

II. Literature Review

Virtual teams operate across geographical and cultural boundaries using digital tools [4]. Their outcomes depend on interconnected factors including communication, technology,

leadership, and trust. This review synthesizes existing literature to establish a theoretical foundation for the Virtual Team Performance Management (VTPM) model, identifying key performance determinants, challenges, and limitations of existing frameworks.

A. Key Determinants of Virtual Team Performance

Communication drives coordination and trust in virtual teams through message clarity, appropriate frequency, open dialogue, and responsive leadership [1], [16], [20], [48]. Mismatches between message medium and task type create problems because complex issues require real-time discussion [38]. Effective communication demands both technical access and relational competence across diverse team backgrounds [18].

Technology management directly impacts performance through tool adequacy, availability, reliability, and user competence [12]. High-performing teams integrate communication, collaboration, and project management systems [25]. Post-deployment training gaps remain common [23], and inadequate tools and information-sharing practices generate both delays and psychological burden in agile virtual teams [24], [32].

Role clarity is essential where informal cues are absent. Ambiguity causes duplicated work, missed tasks, and interpersonal conflict [43]. Defined positions enhance coordination, productivity, and confidence [43], requiring leaders to articulate and revisit responsibilities regularly, particularly in agile software development contexts [15].

Strategic planning in distributed work demands explicit task breakdowns, visible priorities, and established deadlines [22]. Strategic planning encompasses goal setting, activity prioritization, and ensuring members understand their role in the broader project context [47]. Effective planning reduces communication waste; excessive messaging often signals poor planning rather than active collaboration [31].

Documentation preserves organizational memory, retaining decisions and enabling continuity [45]. Quality records enhance transparency and accountability [46], while poor recordkeeping creates operational failures. Performance monitoring, when oriented toward process improvement rather than surveillance, enables early problem identification and works most effectively when integrated with documentation practices [41].

Trust develops through demonstrated reliability and competence [13], requiring structured and consistent management practices in virtual settings where face-to-face interaction is limited [43].

B. Existing Models and Their Limitations

Several models have attempted to address virtual team management, as summarized in Table I.

Despite their contributions, existing models suffer from three common limitations: they lack empirical support, cannot be readily operationalized into actionable management practices [39], or fail to integrate multiple performance determinants simultaneously [22], [27]. Recent studies also highlighted that agile transformation and knowledge management practices influence collaboration, project values, and team coordination within virtual project environments [33], [34].

TABLE I. COMPARATIVE ANALYSIS OF EXISTING VIRTUAL TEAM MANAGEMENT MODELS

Model	Weakness	Comparison / Contribution
Kauffmann & Carmi [23]	Limited attention to collaboration tools and cultural diversity	Emphasizes communication and team integration
Iyamu & Adelakun [22]	Limited practitioner involvement	Focuses more on collaboration than overall team performance
Lumseyfai [27]	Based on a single organizational context	Integrates human interaction, technology, and governance
Holtz [21]	Lacks feedback and leadership communication elements	Highlights engagement and emotional management
Wei [42]	Limited focus on collaboration technologies	Emphasizes team diversity and communication processes
Garro-Abarca [19]	Limited attention to tools and technology management	Highlights communication and trust in team performance

C. Research Gap and Study Contribution

The literature establishes that virtual team performance depends not on geographical distance but on coordinated management practices. Performance is a multidimensional phenomenon requiring integrated attention to communication, technology, planning, documentation, and role clarity. However, no existing model provides an empirically informed, practically operationalizable framework that addresses all these dimensions together.

This study fills this gap by developing the Virtual Team Performance Management (VTPM) model, a systematic, empirically supported framework designed to enhance virtual team performance through integrated management of communication, technology, strategic planning, documentation, and role clarity.

III. METHODOLOGY

This study employs a quantitative research methodology to investigate factors influencing virtual team performance and to develop a structured management model. The methodology is structured into four main phases: theoretical study, empirical study, model development, and model validation.

A. Research Design

A quantitative research design was adopted to allow the organized collection of data and statistical analysis [2]. The four-phase approach consists of: 1) a literature review to determine the issues of interest, 2) an empirical investigation (through a questionnaire- Appendix), 3) the development of the Virtual Team Performance Management Model (VTPM), and 4) expert validation of the model. Fig. 2 illustrates the four-phase research methodology adopted in this study. This is done to make sure that the model is conceptually grounded and quantitatively supported.

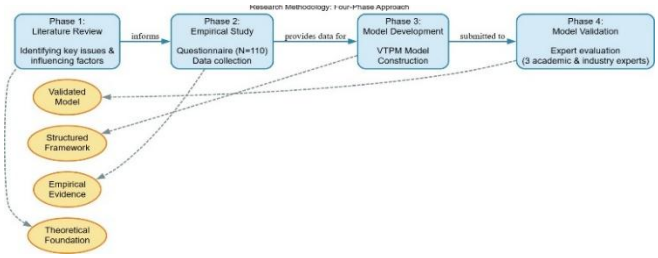


Fig. 2. Four-phase research methodology.

A. Population and Sampling

People with experience in virtual teamwork in different industries were the target population. With the purposive sampling technique, 110 respondents were sampled in the study, and this was considered sufficient to provide a statistical description and a valuable understanding of virtual team practices [43]. The study involved 110 respondents selected using purposive sampling techniques. The sample size was considered sufficient to identify patterns, trends, and perceptions related to factors affecting virtual team performance.

B. Data Collection Instrument

Data were collected via a structured questionnaire developed from the literature review, covering demographics, communication effectiveness, technology management, strategic task planning, role clarity, and documentation practices. All items were measured on a five-point Likert scale to quantify perceptions and support analysis.

C. Instrument Validation and Reliability

The questionnaire was subjected to content validation to ensure that the constructs were measured appropriately and to support reliable statistical analysis of questionnaire data [2], [44]. A pilot study involving 33 respondents was conducted to evaluate the reliability of the research instrument using Cronbach's Alpha analysis through SPSS. The reliability analysis demonstrated acceptable internal consistency across all constructs. The Communication Management construct recorded a Cronbach's Alpha value of 0.829; Technology Management recorded 0.704; Strategic Planning and Effective Task Allocation recorded 0.726; Documentation recorded 0.775; and Role Clarity recorded 0.773. These values exceeded the acceptable threshold of 0.70, indicating satisfactory reliability and suitability of the instrument for quantitative data collection and analysis. Table II presents the reliability analysis results obtained from the pilot study using Cronbach's Alpha.

D. Data Analysis Techniques

The frequency distribution, percentage analysis, and mean score analysis of the virtual team were employed as descriptive statistical tools to identify patterns and examine factors affecting virtual team performance. The analysis focused on identifying patterns, respondent perceptions, and commonly observed factors influencing virtual team performance.

TABLE II. RELIABILITY TEST RESULTS OF PILOT STUDY

Section	Construct	Number of Items	Sample Size	Cronbach's Alpha
A	Demographic Information	13	33	-
B	Communication Management in Virtual Teams	15	33	0.829
C	Technology Management in Virtual Teams	14	33	0.704
D	Strategic Planning and Effective Task Allocation	5	33	0.726
E	Documentation	3	33	0.775
F	Role Clarity	4	33	0.773

E. Model Development

VTPM model was constructed out of the identification of the key factors found in the literature and empirical evidence as core components: communication management, technology management, role clarity, strategic task planning, and documentation. The interdependence of these elements is what constitutes the model structure.

F. Model Validation

Three academic and industry professionals evaluated the relevance, effectiveness, and applicability of the proposed model through expert evaluation. The analysis indicated that the model is realistic and applicable in the practical management of virtual teams.

G. Ethical Considerations

Participation was voluntary, and none had to engage in the research without their consent; they were made aware of the purpose of the research, and all information was kept confidential and was to be used in an academic context only.

H. Summary of Methodology

The methodology offers a systematic and trustworthy way of exploring the performance of a virtual team. The study provides empirical support for the proposed VTPM model since it will be combined with a quantitative analysis and validated with respect to experts.

IV. RESULTS AND DISCUSSION

This section presents the empirical results provided by the questionnaire survey that was conducted among the virtual team practice people. The data are analyzed based on descriptive statistics, which encompass the frequencies, percentages, and means of the data. Table III presents the gender distribution of the respondents involved in the study. The results are organized in the form of major elements that affect Virtual Team Performance Management (VTPM). Table IV summarizes the educational background of the respondents.

TABLE III. GENDER DISTRIBUTION

Gender	Frequency	Percentage (%)
Male	62	56.4
Female	48	43.6
Total	**110**	**100.0**

TABLE IV. EDUCATION LEVEL

Education	Frequency	Percentage (%)
SPM/STPM	2	1.8
Diploma	12	10.9
Bachelor's degree	66	60.0
Postgraduate	29	26.4
PhD	1	0.9
Total	**110**	**100.0**

A. Descriptive Analysis

1) Respondent demographics: A total of 110 respondents participated in this study. Table V presents the types of organizations represented by the respondents.

B. Communication Analysis

The findings reveal that communication is a sensitive issue that influences the performance of virtual teams [19]. Most of the respondents (39.1%) reported the necessity of communication being improved, and the assessment of communication effectiveness has a great discrepancy. The survey indicated that there was inconsistency in leadership communication practices, as 46.4 percent of the surveyed revealed that leaders only communicate on a part-time basis about tasks and roles [15]. Besides, 39.1% of respondents said that they hardly get constructive feedback, implying a significant deficiency in leadership engagement. Table VI summarizes respondent perceptions regarding communication effectiveness in virtual teams.

C. Technology Management Analysis

Most respondents (48.2%) experienced issues occasionally, indicating moderate reliability challenges [42]. Only 35.5% reported sufficient training frequently, indicating a need for improved training strategies [50]. Table VII presents the frequency of technology-related issues experienced in virtual teams.

D. Strategic Planning and Task Allocation

Most respondents agreed that effective strategic planning and task allocation positively influence virtual team performance. Specifically, 56.4% of respondents strongly agreed that strategic planning with effective communication improves productivity within virtual teams, while another 56.4% strongly agreed that poor task allocation and lack of strategic planning result in repetitive and unnecessary communication. Table VIII summarizes respondent perceptions regarding strategic planning and task allocation effectiveness. These findings indicate that structured planning, clear communication, and proper task allocation are essential for improving coordination and overall effectiveness in virtual teams [47].

E. Documentation Analysis

63.6% of respondents agreed that excessive documentation exists, which may reduce efficiency [45].

F. Role Clarity Analysis

Although most respondents agreed that roles are clear, over 50% also reported difficulties when roles are not defined [43].

TABLE V. TYPE OF ORGANIZATION

Organization	Frequency	Percentage (%)
Academic	10	9.1
Public	22	20.0
Corporate	73	66.4
Others	5	4.5
Total	**110**	**100.0**

TABLE VI. COMMUNICATION LEVEL IN VIRTUAL TEAMS

Level	Frequency	Percentage (%)
Very Poor	8	7.3
Poor	10	9.1
Needs Improvement	43	39.1
Satisfactory	39	35.5
Effective	10	9.1
Total	**110**	**100.0**

G. Discussion

The results show that the most important factors in the performance of a virtual team are communication, technology management, and clarity of roles [3], [39]. The issue of communication issues, especially in feedback regarding leadership and clarity of leadership, affects team effectiveness a lot [1], [6]. The issues associated with technology are another cause of inefficiency in that there is a lack of training and problems associated with the tool [12]. Strategic planning enhances productivity when it is correctly applied, but failure to maintain consistent task allocation compromises team performance [29], [47], [49]. Furthermore, too much documentation and role ambiguity create team inefficiencies and confusion [46]. Fig. 3 summarizes the key quantitative findings obtained from the survey respondents.

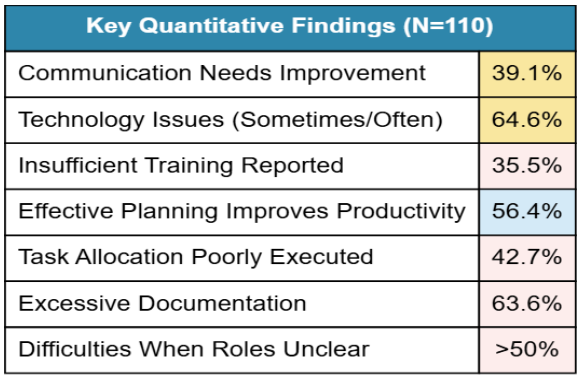


Fig. 3. Key quantitative findings (N=110).

TABLE VII. ISSUES WITH COLLABORATIVE TOOLS

Frequency of Issue	Frequency	Percentage (%)
Never	1	0.9
Rarely	29	26.4
Sometimes	53	48.2
Often	18	16.4
Always	9	8.2
Total	**110**	**100.0**

TABLE VIII. STRATEGIC PLANNING EFFECTIVENESS

Response	Frequency	Percentage (%)
Strongly Disagree	2	1.8
Disagree	8	7.3
Slightly Agree	15	13.6
Agree	23	20.9
Strongly Agree	62	56.4
Total	**110**	**100.0**

In general, the findings can be utilized to work on a structured VTPM model with communication, technology, planning, documentation, and role clarity as the primary aspects of it.

V. VIRTUAL TEAM PERFORMANCE MANAGEMENT (VTPM) MODEL DEVELOPMENT

In this part, the framework of Virtual Team Performance Management (VTPM) is developed. This step is aimed at transforming the findings made during the literature review and the conducted empirical research into a model structure that could be employed to enhance the management and performance of virtual teams. The section covers the logic of this model, the types of factors of the model, a suggested framework, functions of each aspect, and the method of professional validation that is applied to determine the value and applicability of the model. The model was developed following the necessity to offer a practical and holistic structure of organizations and team leaders who work with virtual teams [37]. As the results of the last chapter revealed, various factors constantly recur, as they affect the performance of a virtual team, and the most frequent are the management of communication, the use of technology, planning of the tasks, documenting, and clarity of role in a virtual team. Table IX presents respondent perceptions regarding documentation practices in virtual teams. The findings that were made served as the keystone to determining the key elements of the model and arranging them into a logical framework that helps in performance management in distributed team setups. The influencing factors mentioned above had to be sorted out into categories to build the model beforehand. Table X summarizes the responses related to role clarity within virtual teams. Such a classification narrows down the scope of comprehending the contribution of the various approaches applied by management to the success of virtual teams. The section, accordingly, starts by talking of the classification of the factors and then goes on to describe the proposed model and its validation.

TABLE IX. DOCUMENTATION VOLUME

Response	Frequency	Percentage (%)
Strongly Disagree	3	2.7
Disagree	13	11.8
Slightly Agree	24	21.8
Agree	48	43.6
Strongly Agree	22	20.0
Total	**110**	**100.0**

TABLE X. ROLE CLARITY

Response	Frequency	Percentage (%)
Strongly Disagree	1	0.9
Disagree	12	10.9
Slightly Agree	27	24.5
Agree	44	40.0
Strongly Agree	26	23.6
Total	**110**	**100.0**

A. Categorization of Key Influencing Factors in Virtual Team Performance Management

Classification of the management-related components or elements, based on the empirical findings, was the next step in the model development process. Judging by the questionnaire answers and the discussion in the last section, the most important aspects that influenced the performance of virtual teams were summarized into five major categories:

- Effective Communication Management
- Adequate Technology Management
- Effective Strategic Planning and Task Allocation
- Documentation
- Role Clarity

These five categories were chosen due to their reiteration as the most pertinent and pertinent areas that were relevant in the literature review as well as in the empirical investigation. Their combination constitutes the management aspects that are necessary to coordinate successfully, monitor, and enhance virtual team performance [28]. Mapping the forty-questionnaire metrics into these five categories was also done during the classification process. Most metrics can be found in the communication category since communication was ranked as the most central activity in the operations of a virtual team [1]. Such areas that are covered by technology management are collaboration tools, training, availability of tools, and successful utilization of digital platforms. Strategic planning and task allocation involve the aspects that are related to clear planning, task allocation, and work coordination. Documentation is the process of documenting and maintaining the information regarding the project. The factors addressed by

role clarity are those pertaining to clarity of expectations and responsibility, as well as accountability. The importance of this categorization is that a substantial number of individual survey measures are reduced to a manageable and meaningful structure through this categorization. The model does not consider metrics independently but clumps the related practices into fundamental managerial units. This aids in interpretation as well as practice since the organizations can adopt the model by paying attention to these five big areas rather than solitary concerns.

B. Proposed Virtual Team Performance Management (VTPM) Model

The results of the empirical study conducted in this study were the foundation for the development of the Virtual Team Performance Management (VTPM) model. The suggested model was created to assist organizations and team leaders in controlling their virtual teams more efficiently and enhancing the performance of their teams. The model is grounded on the five most important elements realized in empirical research: Effective Communication Management, Adequate Technology Management, Effective Strategic Planning and Task Allocation, Documentation, and Role Clarity. The components mentioned were based on the results of the surveys and are regarded as crucial in managing virtual teams [30]. Effective Communication Management has been utilized as the go-to mechanism in virtual teams in an effort to ensure easy communication and effective work among team members [20]. Another essential consideration is Adequate Technology Management, which focuses on the utilization of technological resources and platforms, as well as on their oversight and maintenance to support communication and collaboration [42]. In the VTPM model, Effective Strategic Planning and Task Allocation become one of the attributes that are needed to attain alignment with the project objectives, as well as to allocate work efficiently [22]. Documentation is also considered necessary to ensure transparency and that other information and project developments are well recorded and communicated [45]. Role Clarity is another important factor in virtual teams since it makes every team member understand their tasks, role, and the role they have in relation to others in the project as a unified team [43]. The suggested VTPM is a framework of management for enhancing the performance of virtual teams. Fig. 4 presents the proposed Virtual Team Performance Management (VTPM) model developed from the empirical findings. Through taking into account and applying all five primary aspects, organizations and team leaders can better manage the distributed work and have a higher chance of successful completion of the projects.

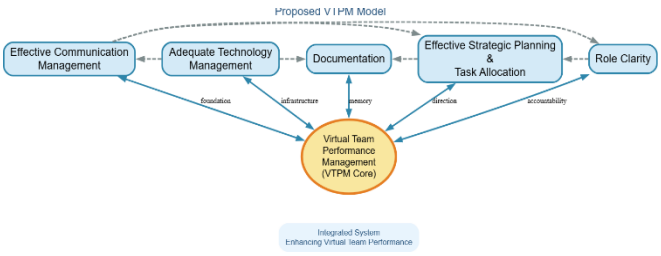


Fig. 4. Proposed Virtual Team Performance Management (VTPM) model.

Fig. 4 above represents the overall structure of the VTPM model. Although the components are presented individually, in practice they are interrelated. Communication influences planning and role clarity, technology supports communication and coordination, documentation reinforces shared understanding, and role clarity improves accountability and productivity. Therefore, the model should be understood as an integrated system rather than a set of isolated management activities.

C. Components of the Virtual Team Performance Management (VTPM) Model

VTPM model includes five components interacting with each other. Effective Communication Management [1], [20] defines ongoing communication as the likeness to collaboration based on the understanding of expectations, active listening, and feedback systems, which should be carried out through the right channels [38]. Teams need appropriate tools and training, so Adequate Technology Management [12], [42] makes sure the teams have the right tools and training in a certain context [25], [32]. Good Strategic Planning and Task Assignment [22], [47] gears project goals and well-defined responsibilities, and frequent reviews of development. One of the ways organizational memory is preserved and transparency is ensured is documentation [45], [46]; however, it has been found to have to be balanced to ensure that it is not inefficient. Lastly, Role Clarity [43] provides both parties, each member, with knowledge of their duties through clear communication of the leadership and continuous support [15]. These elements do not exist in isolation since they are a system. Fig. 5 illustrates the integrated VTPM system and the relationships between its core components.

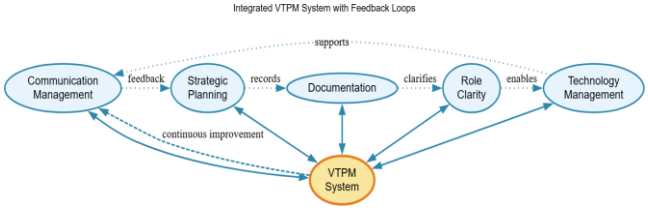


Fig. 5. Integrated VTPM system with feedback loops.

D. Model Validation

1) Introduction: Expert validation is a significant measure that needs to be taken to verify the accuracy, usefulness, and credibility of the VTPM model [2]. This was done through the model being exposed to expert scrutiny and juxtaposition to real-world and hypothetical expectations within the field of virtual team performance management. The validation stage was aimed at the validation of the usability and appropriateness of the VTPM model in the real setting of virtual teams. Validation is significant as a model should not be theoretically sound only, but practical. Domain experts can benefit by providing valuable information about experience and conceptual information. Their analysis allows judging whether the model is reflective of the conditions of real-world virtual teams, whether the model is composed of relevant components, and whether or not enhancements are necessary.

2) Experts involved in the process of validation: The VTPM model was found to be evaluated by academic and industry experts. The academic experts assessed the correspondence of the model to the theoretical and scientific assumptions, and the professionals serving industry assessed the correspondence of the model to the real conditions of work and management in virtual teams [8]. The combination was a balanced and holistic evaluation of the model. The involvement of academic and industrial professionals allowed for defining strengths, weaknesses, and potential changes. This shaped and multi-perspective validation procedure contributed to the strengthening of the credibility of the model and boosted confidence in its applicability in research and practice.

3) Expert validation: The expert validation procedure concentrated on the sub-componential as well as key constituents of the VTPM model. For the Effective Communication Management component, specialists approved the appropriateness of several subcomponents, such as effective communication of the leadership on the task functions and project responsibilities, team-in-the-loop feedback, effective verbal and written communication, strong leadership communication, and the ability to share information using appropriate communication [15]. Further decision making by academic experts regarded these sub-components as adequate, whereas industry experts were broadly in agreement, but were found in a manner the information sharing is practiced.

For the Adequate Technology Management component, the experts confirmed the significance of team leaders working to identify and resolve problems, the necessity of training and education in using collaboration tools, the most efficient utilization of multiple communication and collaboration tools, and better access to these tools [42]. The industry input showed that the use of the tools was proper, but awareness and accessibility could be expanded beyond the scope of tools that were widely used.

For the Role Clarity component, the professionals affirmed that role and responsibility articulation is required and appropriate in the model [43]. They also emphasized cross-functional teamwork and communication, along with mentioning that this process might need additional polishing in large or complicated teams where it is not easy to coordinate interdepartmental work.

For the Effective Strategic Planning and Task Allocation sub-component, academic and industry professionals found the sub-elements appropriate [22]. These involve the implementation of the strategic planning by team leaders through effective communication, task allocation through clear communication, and strategy establishment by the team leaders at the first stage of assigning tasks and responsibilities. Scholars admitted that this element is a powerful contributor to the effectiveness of virtual teams and their fit to the project goals.

For the Documentation component, the specialists ensured the appropriateness of correct documentation goals and document review procedures [45]. These results imply that documentation may be helpful and compatible with the agile or efficient working practices, provided that it is meaningful and controlled properly.

Besides rating sub-components, there was also the rating of the significance of big components by experts. By rating Effective Communication Management, Technology Management, and Documentation as very important, Strategic Planning and Task Allocation as important, and Role Clarity as moderately important, the academic experts rated them accordingly. Technology Management, Strategic Planning and Task Allocation, and Documentation were rated as very important by industry experts, Effective Communication Management important and Role Clarity moderately important by the industry experts. These scores demonstrate that there is high support for the primary structure of the VTPM model.

Professionals also gave recommendations on how it could be improved. Research scholars suggested refinements of the technology management factor in the IT-oriented settings, specialized communication technology to enhance cooperation, and application of the model to teams other than software teams. Experts in the industry recommended smaller, more focused meetings with the help of collaborative tools, more extensive applications of numerous tools in managing time and tasks, improved communication around objectives and key performance indicators, improved evaluation of the status of the existing technology, and organized performance assessment and improvement strategies. These suggestions make the model more useful and flexible. Fig. 6 compares the validation ratings provided by academic and industry experts.

Expert Validation Ratings Comparison		
Component	**Academic Experts**	**Industry Experts**
Communication Management	Very Important	Important
Technology Management	Very Important	Very Important
Strategic Planning	Important	Very Important
Documentation	Very Important	Very Important
Role Clarity	Moderately Important	Moderately Important

Fig. 6. Comparison of academic and industry expert ratings.

VI. Conclusion

This is the concluding section of the study that provides a comprehensive overview of the research done on virtual team performance management, specifically on the communication as well as the technology management strategy employed to manage performance in the virtual project team. This study has synthesized the research objectives, methodology, highlights, and overall findings related to virtual teams operating in the digital age [40]. The research was based on finding a multifocused explanation of how the proposed Virtual Team Performance Management (VTPM) model can solve the challenges determined in the virtual team setting. This study used both theoretical models combined with the empirical evidence presented in real-life situations to create a strong model of virtual team performance optimization in different situations. This section emphasizes that the study not only enhances the theoretical base of virtual team management but

also offers practical guidance on how it can be applied. The importance of these findings does not support just academic discourse, but provides useful information to virtual team leaders, team members, and the organization that deals with the management of virtual projects. This study is able to present actionable suggestions and strategic considerations to serve as guidance to stakeholders to resolve the problem of virtual collaboration. In general, the section is a valuable source in relation to the future of the developing virtual team's scenario.

A. Analysis of the Research Results

This part is a summary of the study progress according to the three research objectives. The initial goal, which was to determine problems and difficulties encountered by virtual teams in performance management, was realized in terms of the literature review and questionnaire results. As it was found, the key challenges are communication ambiguity, poor application of technology, documentation weaknesses, deficient strategic planning, poor allocation of tasks, as well as role ambiguity [22]. These problems formed a good background for the realization of the necessity to have a structured and efficient management strategy. Fig. 7 summarizes the research objectives and the corresponding achievements of the study.

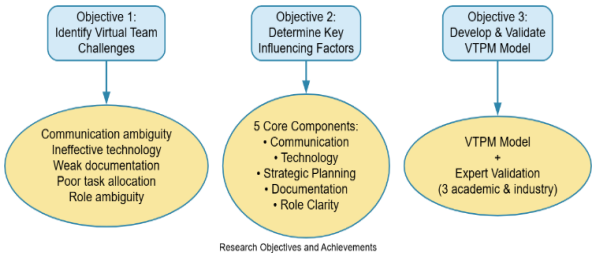


Fig. 7. Research objectives and key achievements.

The second goal, aimed at defining the main factors that affect the virtual team performance, was fulfilled in selecting five fundamental elements consisting of communication management, technology management, strategic planning, task allocation, role clarity, and documentation. All these aspects were based on empirical evidence and attested by the existing literature [3], [39]. They constitute the foundation for defining major performance criteria that should be dealt with by organizations and team leaders to enhance the performance of virtual teams. The third goal that attempted to formulate the Virtual Team Performance Management (VTPM) model was met with great success through the use of empirical evidence combined with theoretical methodologies and expert guidance [28], [37]. The VTPM model developed not only solves the major problems discussed but also offers an opportunity to have a practical design with the best practices, evidence-based management approaches, and feedback systems. The model was further refined through expert evaluation in order to make sure that it matches the project management requirements in the real world [8]. As a whole, this development of the study fits well within the three research objectives. The research identifies and evaluates the factors affecting virtual team performance and develops a practical management model.

B. Significance of the Study

The research on the Virtual Team Performance Management (VTPM) model makes a great contribution to both theoretical, practical, and social levels. Fig. 8 illustrates the theoretical, practical, and societal contributions of the proposed VTPM model.

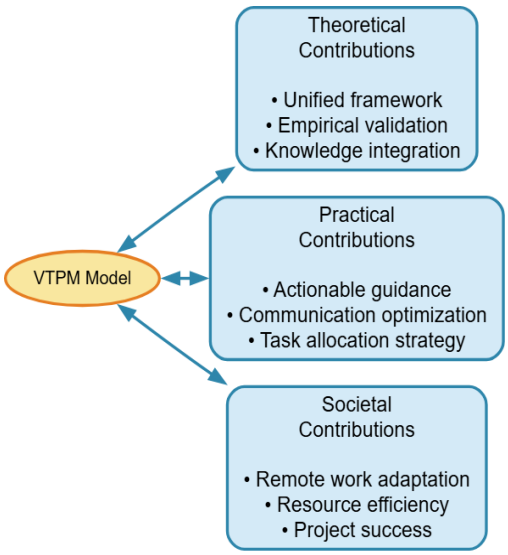


Fig. 8. VTPM model contributions: theoretical, practical, and societal.

1) Theoretical contributions: The study contributes to the theoretical knowledge on the issue of virtual teams by uniting the main factors that include management of communication, management of technology, strategic planning, task division, clarity of tasks, and documentation. VTPM model is an overall framework that integrates all these aspects in a single framework for managing virtual teams [26], [30]. Moreover, the empirical findings obtained through surveys and experts' evaluation give rather good evidence on the effectiveness of the model. This helps in closing the gap between theory and practice as it provides a validated framework to be used in further research.

2) Practical contributions: In practical terms, the VTPM model has implications for the practical understanding of team leaders and organizations. It provides systematic leadership in ensuring effectiveness in communication and technology use, as well as a better strategy and delegation of tasks [14], [47]. The model can also greatly enhance the performance and the productivity of the team in a virtual environment because some typical challenges, including barriers to communication, knowledge deficits, and poor management of technology, have been solved.

3) Societal contributions: The analysis aids the organizations in readjusting to the growing trends of a remote and distributed workplace setting [9]. The VTPM model helps enhance more fruitful and fulfilling remote working circumstances by enhancing the management of virtual teams. Moreover, the study should minimize instances of project failure as a result of communication and technology problems, which helps to ensure a higher level of efficiency in the use of resources and a better result of the project at the organizational level.

C. Critical Limitations and Threats to Validity

This study has several limitations that qualify its conclusions:

1) Cross-sectional design: Data were collected at one point; causal relationships cannot be inferred. Longitudinal studies are needed.

2) Self-report bias: All measures are perceptual. Objective performance metrics (e.g., project completion time, deliverable quality) were not collected.

3) Sample size and generalizability: N=110 is moderate. The sample is predominantly Malaysian corporate employees (66.4%); findings may not generalize to other cultural contexts, very small teams (<5 members), or very large teams (>20 members).

4) Lack of objective performance data: Perceived team performance may not align with actual outcomes.

5) Expert validation sample: Only three experts; a larger panel would increase confidence.

6) No structural equation modeling (SEM): Mediation pathways (e.g., whether communication effects are mediated by trust) were not tested.

7) Scope constraints: The VTPM model is validated only for technology-mediated, task-oriented teams in stable connectivity environments. It may not apply to emergency-response, creative, or purely social virtual teams.

D. Future Research Recommendations

It is suggested that future studies be done to enhance the VTPM model, especially as regards virtual and agile teams. These are the recommendations that fall into three broad regions.

1) Virtual team performance measurement: Information shedding in case of future research needs to involve detailed case research in an organization where virtual teams are significantly applied. This would give more detailed information on how the performance of teams is assessed in a real-life setting [10]. The most effective way to gather data on the effectiveness of the VTPM model is by means of documenting the project, interactions between the team, and interviews with the team members and managers. Longitudinal studies can also be performed in order to monitor the change in performance over time.

2) Formulation of performance measurement metrics: the future study must be on how to come up with developed performance measures in a well-structured, objective manner. Suggested metrics include: Sprint burndown, cycle time to assess productivity [41]. Division of high-quality deliverables to determine the output quality. Minutes during which meetings were undertaken to assess the efficiency of communication [38]. The metrics can be used to complement the current performance assessment models and offer more extensive information applicable to the management of virtual teams.

3) Pay attention to agile virtual teams: Future studies should address the implications of the VTPM model to Agile Virtual Teams and software development environments within the digital economy [35], [17], [36]. The research established that performance appraisal among agile virtual teams has not been extensively studied, and mostly it is based on subjective appraisals [47]. The way VTPM elements can accommodate agile methodologies like Scrum and Kanban should be studied in the future [50]. Moreover, innovative analytics and machine learning methods can be considered to forecast the performance of the team and suggest changes based on the information in real time.

Summing it up, the presented study produces a holistic approach to enhancing the performance of virtual teams using the Virtual Team Performance Management (VTPM) model. The model touches upon such essential issues as communication management, technology management, strategic planning, task distribution, role clarity, and documentation. The model may incorporate potential performance evaluation metrics such as cycle time, burndown charts of sprints, and quality indicators, among others, which makes it a solid framework and tool that can assess and enhance team effectiveness. Time tracking of the meetings is also helpful in communication and cooperation. VTPM model provides a bridge between theory and practice because it provides practical recommendations on how to manage virtual teams. It allows ongoing improvement, decision-making, and resource allocation, which will increase the success of a particular project. Further studies on the model should be done in the future by incorporating detailed performance measurement models and sophisticated methods of analysis. It will help to shed more light on virtual team dynamics and facilitate decision-making based on the facts. On balance, the improved VTPM model could help to advance the practice of virtual team management considerably and enhance the sphere of virtual project management, in general.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Universiti Kebangsaan Malaysia for supporting this research.

REFERENCES

[1] S. Agarwal, S. Ferdousi, T. Stahl, M. John, and A. Nalven, "Effective leadership in virtual teams during the COVID-19 pandemic," *Engineering and Technology Management Student Projects*, 2020.

[2] I. Ahmed and S. Ishtiaq, "Reliability and validity: Importance in medical research," *Journal of the Pakistan Medical Association*, vol. 71, no. 10, pp. 2401–2406, 2021.

[3] H. M. A. Al Kaabi, S. Sidek, and N. A. Mosali, "Mediation model of factors affecting virtual team performance in UAE organisations," *International Journal of Sustainable Construction Engineering and Technology*, vol. 13, no. 2, pp. 268–284, 2022.

[4] A. Alaiad, Y. Alnsour, and M. Alsharo, "Virtual teams: Thematic taxonomy, constructs model, and future research directions," *IEEE Transactions on Professional Communication*, vol. 62, no. 3, pp. 211–238, 2019.

[5] A. B. Anoye and J. S. Kouamé, "Leadership challenges in virtual team environment," *International Journal of Scientific and Technology Research*, vol. 7, no. 7, pp. 160–167, 2018.

[6] C. C. S. de Araújo, C. D. Pedron, and F. Q. P. de O. e Silva, "IT project manager competencies and team commitment: A new scale proposal," *Revista de Gestão e Projetos*, vol. 9, no. 1, pp. 39–57, 2018.

[7] P. Barry and B. Kane, "Global virtual team working during the COVID-19 pandemic," *Interacting with Computers*, vol. 35, no. 5, pp. 681–690, 2023.

[8] C. J. Behling, "Project success in virtual projects: A qualitative study of leadership behaviors," Ph.D. dissertation, 2019.

[9] S. K. Bhat, N. Pande, and V. Ahuja, "Virtual team effectiveness: An empirical study using SEM," *Procedia Computer Science*, vol. 122, pp. 33–41, 2017.

[10] M. S. Bissaliyev, "The effectiveness of collaboration tools on virtual project management," *International Journal of Applied Engineering Research*, vol. 12, no. 21, pp. 10747–10755, 2017.

[11] C. Breuer, J. Hüffmeier, F. Hibben, and G. Hertel, "Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams," *Human Relations*, vol. 73, no. 1, pp. 3–34, 2020.

[12] V. Casey, "Virtual software team project management," *Journal of the* Brazilian *Computer Society*, vol. 16, no. 2, pp. 83–96, 2010.

[13] P. Chamakiotis, N. Panteli, and R. M. Davison, "Reimagining e-leadership for reconfigured virtual teams due to COVID-19," *International Journal of Information Management*, vol. 60, p. 102381, 2021.

[14] D. Ciric, B. Lalic, D. Gracanin, I. Palcic, and N. Zivlak, "Agile project management in new product development and innovation processes: Challenges and benefits beyond software domain," in *TEMSISIE 2018 – 1st Annual International Symposium on Innovation and Entrepreneurship of the IEEE Technology and Engineering Management Society*, pp. 1–9, 2018.

[15] K. A. Crowne, "Does national culture influence peer evaluations on global virtual teams?" *Journal of Teaching in International Business*, vol. 31, no. 3, pp. 191–213, 2020.

[16] I. Dumitras¸cu-Ba˘lda˘u and D. D. Dumitras¸cu, "Intercultural communication and its challenges within the international virtual project team," *MATEC Web of Conferences*, vol. 290, 2019.

[17] I. Dumitras¸cu-Ba˘lda˘u and O. Dumitras¸cu, "Research on the behavior of factors that influence the international virtual project team performance, using data modeling techniques," *Sustainability*, vol. 11, no. 3, 2019.

[18] J. S. Gallego, I. Ortiz-Marcos, and J. Romero Ruiz, "Main challenges during project planning when working with virtual teams," *Technological Forecasting and Social Change*, vol. 162, p. 120353, 2021.

[19] V. Garro-Abarca, P. Palos-Sanchez, and M. Aguayo-Camacho, "Virtual teams in times of pandemic: Factors that influence performance," *Frontiers in Psychology*, vol. 12, pp. 1–14, 2021.

[20] C. Hargreaves, A. P. Clarke, and K. R. Lester, "Microsoft Teams and team performance in the COVID-19 pandemic within an NHS trust community service in North-West England," *Team Performance Management*, vol. 28, no. 1–2, pp. 79–94, 2022.

[21] K. Holtz, V. O. Castella, A. Z. Abad, and B. Gonza´lez-Anta, "Virtual team functioning: Modeling the affective and cognitive effects of an emotional management intervention," *Group Dynamics*, vol. 24, no. 3, pp. 153–167, 2020.

[22] T. Iyamu and O. Adelakun, "A global virtual team model to improve software development collaboration project," *Information Systems and e-Business Management*, vol. 19, no. 3, pp. 937–956, 2021.

[23] D. Kauffmann and G. Carmi, "A comparative study of temporary and ongoing teams on e-environment," *IEEE Transactions on Professional Communication*, vol. 62, no. 2, pp. 148–164, 2019.

[24] F. Kortum, J. Klunder, O. Karras, W. Brunotte, and K. Schneider, "Which information help agile teams the most? An experience report on the problems and needs," in *46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 306–313, 2020.

[25] C. T. Kuah and K. Y. Wong, "Evaluating team performance using slack based data envelopment analysis," in *International Conference on Advanced Communication Technology (ICACT)*, pp. 1152–1157, 2014.

[26] P. Lous, P. Tell, C. B. Michelsen, Y. Dittrich, and A. Ebdrup, "From scrum to agile: A journey to tackle the challenges of distributed development in an agile team," in *ACM International Conference Proceeding Series*, pp. 11–20, 2018.

[27] J. Lumseyfai, "A four-pillared holistic model for improving performance in engineering virtual project teams," *Engineering Management Journal*, vol. 32, no. 2, pp. 107–119, 2020.

[28] J. Malmqvist, K. Hellberg, G. Möllås, R. Rose, and M. Shevlin, "Conducting the pilot study: A neglected part of the research process? Methodological findings supporting the importance of piloting in qualitative research studies," *International Journal of Qualitative Methods*, vol. 18, pp. 1–11, 2019.

[29] A. Mashmool, S. Khosravi, J. H. Joloudari, I. Inayat, T. J. Gandomani, and A. Mosavi, "A statistical model to assess the team's productivity in agile software teams," in *CANDO-EPE 2021 – IEEE 4th International Conference and Workshop in Obuda on Electrical and Power Engineering*, pp. 11–18, 2021.

[30] C. Mayer, S. Mutze-Niewo¨hner, and V. Nitsch, "Empirical classification of advanced information technology towards their support of leadership behaviors in virtual project management settings," in *IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 275–279, 2020.

[31] K. Mehtab, A. ur Rehman, S. Ishfaq, and R. A. Jamil, "Virtual leadership: A review paper," *Mediterranean Journal of Social Sciences*, vol. 8, no. 4–1, pp. 183–193, 2018.

[32] S. Morrison-Smith and J. Ruiz, "Challenges and barriers in virtual teams: A literature review," *SN Applied Sciences*, vol. 2, no. 6, pp. 1–33, 2020.

[33] P. Paterek, "Agile transformation in project organization: Knowledge management aspects and challenges," in *Proceedings of the European Conference on Knowledge Management (ECKM)*, vol. 2, pp. 1170–1179, 2017.

[34] P. Paterek, "Agile transformation changes from the perspective of project team values," in *8th International Scientific Conference on Project Management in the Baltic Countries: Project Management Development – Practice and Perspectives*, pp. 162–174, 2019.

[35] K. Saeedi and A. Visvizi, "Software development methodologies, HEIs, and the digital economy," *Education Sciences*, vol. 11, no. 2, pp. 1–22, 2021.

[36] B. Seerat, M. Samad, and M. Abbas, "Software project management in virtual teams," in *Proceedings of the Science and Information Conference (SAI)*, pp. 139–143, 2013.

[37] D. Simpson, "Advantages and disadvantages of international virtual project teams," *International Business and Global Economy*, vol. 36, no. 1, 2017.

[38] V. Stray, N. B. Moe, and M. Noroozi, "Slack me if you can! Using enterprise social networking tools in virtual agile teams," in *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)*, pp. 111–121, 2019.

[39] C. K. Tan, T. Ramayah, A. P. Teoh, and J. H. Cheah, "Factors influencing virtual team performance in Malaysia," *Kybernetes*, vol. 48, no. 9, pp. 2065–2092, 2019.

[40] N. U. Thuong, "Virtual teams in the digital age," *Hue University Journal of Science: Economics and Development*, vol. 128, no. 5C, pp. 67–80, 2019.

[41] M. Topaloglu and A. S. Anac, "Exploring major factors affecting virtual team performance," *European Journal of Business and Management Research*, vol. 6, no. 5, pp. 107–114, 2021.

[42] L. H. Wei, R. Thurasamy, and S. Popa, "Managing virtual teams for open innovation in Global Business Services industry," *Management Decision*, vol. 56, no. 3, pp. 570–590, 2018.

[43] S. I. Wong and S. van Gils, "Initiated and received task interdependence and distributed team performance: The mediating roles of different forms of role clarity," *AI and Society*, 2021.

[44] K. Y. Yigzaw, A. Michalas, and J. G. Bellika, "Secure and scalable statistical computation of questionnaire data in R," *IEEE Access*, vol. 4, pp. 4635–4645, 2016.

[45] K. Yoon and Y. Zhu, "Social media affordances and transactive memory systems in virtual teams," Management Communication Quarterly, vol. 36, no. 2, pp. 235–260, 2021.

[46] W. Yu, K. Cormican, Q. Wu, and S. Sampaio, "In whom do we trust? Critical success factors impacting intercultural communication in multicultural project teams," International Journal of Information Systems and Project Management, 2021.

[47] A. M. Younus and H. Younis, "Conceptual framework of agile project management affecting project performance: Key requirements and challenges," International Journal of Innovative Research in Engineering & Management, vol. 8, no. 4, pp. 10–14, 2021.

[48] J. H. Yahaya, M. M. Basir, and A. Deraman, "Unified communication and collaboration model for virtual distributed teamwork: A study in Malaysia," International Journal of Software Engineering and Its Applications, vol. 9, no. 2, pp. 125–142, 2015.

[49] Z. Ahmed, Z. Mansor, K. Ahmad, and M. Zafarullah, "Enhancing team-related knowledge sharing in agile virtual teams: Social, psychological and trust enablers for distributed collaboration – KSAVT framework," TPM – Testing, Psychometrics, Methodology in Applied Psychology, vol. 32, no. S9, pp. 680–683, 2025.

[50] L. Moradi, Y. Yahya, and I. Mohamed, "Relationship between E-training in virtual team and IT project performance with the mediation role of organizational commitment in E-tourism," in Proc. IEEE, 2017, pp. 1–5.

APPENDIX

Table XI provides the mapping between each construct, the corresponding questionnaire items, and their sources. All items were measured on a 5-point Likert scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree).

The full questionnaire consisted of 22 items across the five independent constructs and multiple items for the dependent variable (team performance). The complete instrument is available from the corresponding author upon request. Table XI presents the mapping of constructs, questionnaire items, and supporting references used in the study.

TABLE XI. CONSTRUCT MAPPING AND QUESTIONNAIRE ITEMS

Construct	Sample Item	Source
Communication	"Leadership clearly communicates task expectations"	[1], [20]
Technology Management	"Collaboration tools are reliable and accessible"	[12], [42]
Strategic Planning	"Project goals are clearly defined at the outset"	[22], [47]
Role Clarity	"My responsibilities are well-defined"	[43], [15]
Documentation	"Documentation is excessive and reduces efficiency" (reverse coded)	[45], [46]
Team Performance (DV)	"Overall, my virtual team performs effectively"	[10]

An Explainable XGBoost-Based Framework for Robust Multi-Cohort Prediction of Pancreatic Cancer

Nada Ahmed El-Gammal[1]*, Rania Ahmed Abdel Azeem Abul Seoud[2], Sayed T. Muhammad[3]
Department of Computers and Systems Engineering-Faculty of Engineering, Fayoum University, Fayoum, Egypt[1, 3]
Department of Electrical Engineering-Specialization in Communications and Electronics Engineering-Faculty of Engineering, Fayoum University, Fayoum, Egypt[2]

Abstract*—Pancreatic cancer remains a leading cause of cancer-related mortality due to its asymptomatic progression and late-stage diagnosis. Early detection is critical for improving patient prognosis and clinical outcomes. Traditional diagnostic approaches and previous computational models often struggle with molecular heterogeneity and technical variations across different genomic platforms. These batch effects limit the reliability and generalizability of predictive biomarkers when applied to diverse clinical settings. This research proposes a robust machine learning framework designed for platform-invariant pancreatic cancer prediction. Large-scale transcriptomic datasets, including microarray data from the Gene Expression Omnibus (GEO) and RNA-seq data from The Cancer Genome Atlas (TCGA), were integrated. Subsequently, the ComBat algorithm was applied to correct batch effects. This resulted in a discovery cohort of 441 samples and an external validation set of 409 samples. An optimized XGBoost classifier was developed through comparative benchmarking. It was compared against several learners, including Random Forest, LightGBM, Support Vector Machines (SVM), and Logistic Regression. The model demonstrated high predictive performance, achieving an internal test AUC of 0.923. External validation was performed across six independent cohorts, yielding a mean AUC of 0.761 ± 0.090 (95% CI: 0.689–0.833). These findings support the robustness and cross-platform generalizability of the proposed framework. To enhance model interpretability, SHapley Additive exPlanations (SHAP) analysis was employed to identify key molecular drivers. These drivers were further validated using biological enrichment analysis through Over-Representation Analysis (ORA) and log2FC-weighted Gene Set Enrichment Analysis (GSEA). The proposed framework provides a reliable and scalable solution for multi-platform integration. This approach facilitates accurate risk stratification and precision oncology in clinical practice.

Keywords*—Pancreatic cancer; gene expression analysis; XGBoost; SHAP explainability; pathway enrichment; Explainable AI (XAI)

I. Introduction

Pancreatic cancer remains one of the most lethal malignancies worldwide, with a five-year survival rate below 10% [1],[14]. This poor prognosis stems from late-stage diagnosis, as the disease is often asymptomatic initially. Accordingly, most patients are being diagnosed only at advanced stages [2]. Since current imaging and biomarkers lack sufficient sensitivity, there is an urgent need for robust molecular signatures to support early diagnostic precision [18].

High-throughput transcriptomic profiling enables comprehensive characterization of pancreatic tumorigenesis [11],[12]. While these datasets offer opportunities to identify molecular signatures, traditional statistical methods struggle with high-dimensional omics data [15]. Machine learning (ML) has improved predictive performance. However, technical variability and a lack of interpretability still limit its clinical applicability.

Existing ML models for pancreatic cancer often suffer from small sample sizes and lack robust cross-cohort validation [16],[18]. Furthermore, many architectures function as black boxes, offering minimal biological insight into their predictive logic. This absence of functional validation hinders the transition from computational models to clinical utility.

Moreover, prior harmonization frameworks frequently focused on isolated objectives, such as batch correction or classifier optimization. Biological interpretability and external validation were often not integrated into a unified analytical pipeline. Consequently, many existing approaches remained vulnerable to platform dependency and limited generalizability across heterogeneous transcriptomic cohorts.

To address these limitations, this study proposed an integrated machine learning framework for robust pancreatic cancer prediction. First, a unified transcriptomic baseline was established by integrating microarray and RNA-seq datasets using ComBat-based batch effect correction. This ensured cross-platform consistency. Second, five machine learning algorithms were systematically benchmarked. This process yielded an optimized XGBoost classifier capable of capturing complex nonlinear gene interactions with superior sensitivity [17]. Third, SHAP analysis was employed to improve clinical interpretability by identifying key molecular drivers underlying model predictions. Fourth, biological relevance was reinforced through integrated functional enrichment analysis using GSEA and ORA. Finally, robustness and clinical reliability were validated across six independent external cohorts. These results supported real-world generalizability.

The remainder of this study is organized as follows. Section II reviews related work on pancreatic cancer prediction. Section III describes the materials and methods, including data integration, batch correction, and framework development. Section IV presents the evaluation criteria. Section V reports the experimental results and performance analysis. Section VI discusses the clinical and biological

*Corresponding author.

implications of the findings. Finally, Section VII concludes the study and outlines future research directions.

II. Related Work

The computational landscape for pancreatic cancer (PC) diagnosis has transitioned from simple classification on small datasets to complex, multi-cohort integrative frameworks. Current literature can be segmented into three primary domains: single-platform diagnostic models, multi-omics integration with batch-effect mitigation, and explainable AI (XAI) for clinical biomarker discovery.

Initial efforts in PC diagnosis often focused on localized cohorts. For instance, Carrillo-Perez et al. [6] utilized machine learning to identify a 6-gene signature in peripheral blood mononuclear cells for PC diagnosis. Their study demonstrates high accuracy in distinguishing PC from chronic pancreatitis. However, it remains centered on transcriptomic data from specific blood samples. This highlights a common trend in single-platform studies where the diagnostic potential is robust within the chosen medium. However, additional cross-platform validation is often required to ensure generalizability across diverse clinical settings, such as tissue-based RNA-seq or microarray datasets.

To overcome the limitations of individual classifiers, recent studies have pivoted toward ensemble architectures. Rao and Prasad [19] developed an ensemble learning model for the classification of gene expression from RNA-seq data for pancreatic cancer prognosis. Their research emphasizes that combining multiple learners significantly reduces the risk of false results common in classical analysis. However, while their approach demonstrates the efficacy of ensemble methods in prognosis, there remains a need for systematic benchmarking that specifically optimizes XGBoost within a framework designed for cross-platform robustness. This ensured high predictive performance was maintained when the model was transitioned from RNA-seq to microarray environments.

As the field moves toward large-scale integration, handling technical noise becomes paramount. Ge et al. [23] addressed this by integrating transcriptomic, methylation, and mutational data across 13 independent cohorts. Their study successfully identified molecular subtypes and the role of the A2ML1 gene. They employed extensive multi-omics integration. However, such large-scale fusions often face significant batch effects between different genomic platforms. This necessitates the use of explicit correction methods to ensure that the identified subtypes reflect true biological heterogeneity rather than platform-specific artifacts.

To bridge the gap between model performance and clinical trust, recent studies have incorporated interpretability tools. Almisned et al. [5] developed an ensemble voting classifier for PC diagnosis and employed SHAP to interpret the model's clinical attributes. Their work provides a clear link between features and predictions. However, a persistent gap remains in the integration of these SHAP-driven insights with formal biological pathway validation (such as GSEA or ORA). Without this multi-layered validation, the clinical translation of black-box models remains a challenge.

Despite strong internal performance, many previous frameworks remained sensitive to platform variability and transcriptomic heterogeneity. As a result, models trained on single-cohort or single-platform datasets often showed limited cross-platform transferability. To address these limitations, the study developed a platform-invariant framework integrating ComBat-based harmonization, benchmarked XGBoost optimization, and SHAP-driven biological validation. External validation across six independent cohorts further supported the framework's clinical robustness and biological relevance.

III. Materials and Methods

The systematic workflow of this study is organized into four phases (Fig. 1). Phase 1 involves multi-platform data integration and batch effect correction using the ComBat algorithm. In Phase 2, five machine learning classifiers are benchmarked, with XGBoost selected for its superior sensitivity and PR-AUC. Phase 3 utilizes SHAP-based explainability to identify key features and validate their biological relevance through ORA and GSEA. Finally, Phase 4 confirms the model's generalizability via independent validation across six unseen cohorts to establish a reliable diagnostic signature.

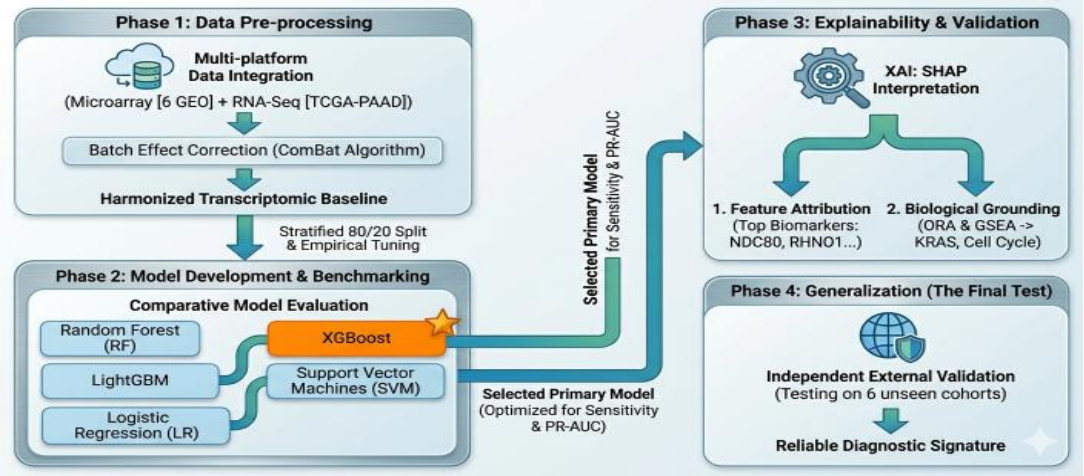


Fig. 1. Overview of the proposed predictive pipeline.

A. Data Collection and Study Design

Publicly available transcriptomic datasets were curated from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) [11]. Multiple microarray cohorts representing pancreatic tumor and normal pancreatic tissue samples were selected. To ensure cross-platform robustness, both microarray and RNA-seq data were integrated. A total of 623 samples were initially collected across all datasets. However, after stringent quality control and the removal of samples with incomplete metadata, 441 high-quality samples remained for training. Gene-level mapping was performed to resolve platform-specific nomenclature, resulting in a unified feature space of 18,734 common genes. This yielded a final training matrix of 441 samples × 18,734 genes for model development. The TCGA pancreatic adenocarcinoma cohort was used as a primary RNA-seq dataset [11], while GEO cohorts were used as the primary microarray datasets. Detailed cohort distributions and platform specifications (tumor-normal sample distribution) are summarized in Table I.

TABLE I. DATASETS USED FOR MODEL DEVELOPMENT

Dataset	Source	Platform	Genes	T/N	All	Role
GSE22780	GEO	Microarray (GPL570)	23,520	8 / 8	16	Discovery
GSE71989	GEO	Microarray (GPL570)	23,520	14 / 8	22	Discovery
GSE62165	GEO	Microarray (GPL13667)	20,034	118 / 13	131	Discovery
GSE28735	GEO	Microarray (GPL6244)	18,883	45 / 45	90	Discovery
GSE62452	GEO	Microarray (GPL6244)	18,883	69 / 61	130	Discovery
GSE16515	GEO	Microarray (GPL570)	23,520	36 / 16	52	Discovery
TCGA-PAAD	TCGA	RNA-seq	60,660	178 / 4	182	Discovery
GSE19279	GEO	Microarray (GPL96)	13515	9 / 6	15	External Validation
GSE15471	GEO	Microarray (GPL570)	23520	39 / 39	78	External Validation
GSE91035	GEO	Microarray (GPL22763)	21189	25 / 25	50	External Validation
GSE60980	GEO	Microarray (GPL14550)	22070	164 / 18	182	External Validation
GSE43795	GEO	Microarray (GPL10558)	31334	26 / 5	31	External Validation
GSE55643	GEO	Microarray (GPL6480)	19595	45 / 8	53	External Validation

B. Gene Annotation and Data Harmonization

The raw expression matrices initially comprised 60,660 Ensembl identifiers. These identifiers were mapped to official HGNC gene symbols via the MyGene.info API. Before querying, Ensembl version suffixes were stripped to maximize mapping efficiency. In cases of many-to-one mappings, expression values were collapsed by calculating the mean. To ensure feature consistency across diverse platforms, including Affymetrix microarray and Illumina RNA-seq datasets, 18,734 common genes were retained. These genes were consistently present across the majority of cohorts.

To ensure comparability, the integrated discovery set was subjected to log2 transformation and quantile normalization. KNN imputation was then applied (k=5) to estimate missing values. Technical heterogeneity was then systematically corrected using the neuroCombat empirical Bayes framework [3],[4]. The dataset source was designated as the batch variable while protecting the biological signal (tumor vs. normal) as a covariate. This batch correction was applied strictly to the discovery dataset preceding the train-test split. Crucially, the six external cohorts were processed independently and excluded from harmonization. This ensured a rigorous evaluation of unseen clinical data. Finally, Z-score standardization was applied to the corrected feature space. This established a robust foundation for both model training and feature attribution using SHAP.

C. Predictive Model Development and Benchmarking

To identify the most robust predictive architecture, five machine learning models [19] were benchmarked, including XGBoost, Random Forest, LightGBM, SVM, and Logistic Regression [5],[7]. All models were evaluated using a stratified 80/20 train–test split. For reproducibility, XGBoost hyperparameters were tuned using grid search on the training cohort. The search space included learning rates (0.001, 0.01, 0.1), maximum depth (3, 5, 7), and regularization terms ($\gamma \in [0, 0.5]$, $\lambda \in [1, 3]$, $\alpha \in [0, 2]$). The optimal configuration was a max_depth of 3, learning rate of 0.01, gamma=0.3, lambda=2, and alpha=1. To mitigate overfitting, subsampling and colsample_bytree were both set to 0.7. Early stopping with 50 rounds was also used on a validation set. The classification threshold was optimized using the Youden Index to improve diagnostic performance. Based on internal benchmarking, XGBoost was selected for interpretation and external validation. It showed the best overall performance. It achieved high sensitivity (0.881) and PR-AUC (0.962).

D. Model Evaluation and External Validation

Performance was initially benchmarked across all five candidate models using AUC-ROC, F1-score, Matthews Correlation Coefficient (MCC), sensitivity, and specificity. Following internal benchmarking, the top-performing XGBoost model was subjected to a rigorous external validation phase. The model was deployed on six additional, independent GEO cohorts. These cohorts are GSE15471, GSE55643, GSE19279, GSE43795, GSE60980, and GSE91035, totaling 409 samples. These validation datasets were kept entirely unseen during initial integration and training stages. This serves as a definitive test of the model's cross-platform robustness.

E. Model Explainability and Biological Interpretation

To move beyond black-box predictions, SHapley Additive exPlanations (SHAP) were employed, specifically for the optimized XGBoost classifier [7]. SHAP was applied to quantify the contribution of individual genes to the model's output [8],[13]. Global feature importance was derived by calculating the mean absolute SHAP values across the test set, identifying the key drivers of the classifier's decisions.

To ensure these features reflected meaningful biology rather than technical noise, a multi-layered validation pipeline was implemented. First, Over-Representation Analysis (ORA) was performed on the annotated top-ranked SHAP genes to identify enriched Gene Ontology (GO) terms [10],[25].

Subsequently, GSEA was conducted using log2FC-weighted p-values to rank genes across Hallmark and KEGG collections [9]. This analysis confirmed that the identified predictors were biologically associated with oncogenic pathways, including cell-cycle dysregulation and KRAS signaling.

IV. Evaluation Criteria

To address class imbalance across internal and external cohorts, the evaluation looked beyond accuracy to analyze PR-AUC, F1-score, and MCC. Sensitivity was strictly prioritized to minimize false negatives that risk fatal delays in clinical intervention.

The following metrics were utilized to quantify the model's performance:

- Accuracy: Measures the overall proportion of correct predictions (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

- Sensitivity (Recall): Measures the ability to correctly identify PC cases (crucial for early detection) (2).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

- Precision: Represents the proportion of positive identifications that were actually correct (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

- F1-Score: The harmonic mean of Precision and Sensitivity, balancing the two metrics (4).

$$\text{F1 score} = 2\frac{\text{Precision.Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

- Matthews Correlation Coefficient (MCC): Provides a balanced measure of quality, even when classes are of very different sizes (5).

$$\text{MCC} = \frac{\text{TP.TN} - \text{FP.FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (5)$$

- AUC-ROC: Measures the model's ability to distinguish between classes by integrating the True Positive Rate (TPR) against the False Positive Rate (FPR) across all thresholds (6).

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(\text{t}))\, \text{dt} \quad (6)$$

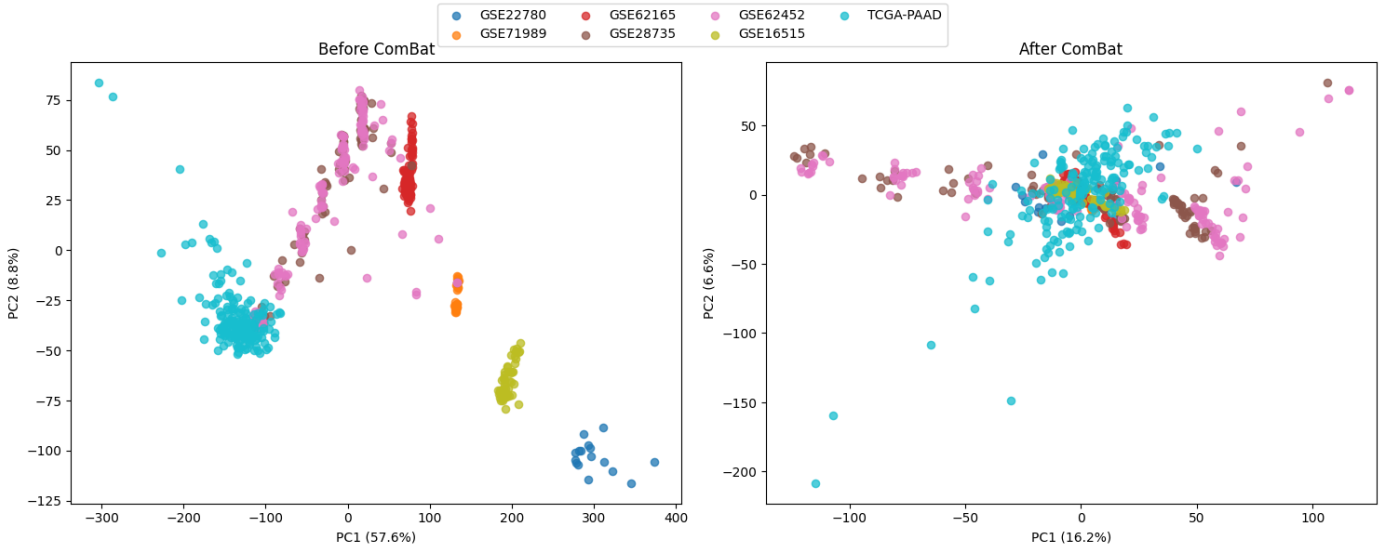
V. Results

The results of this study are presented in five stages:

- Harmonization: Correcting batch effects to integrate multi-platform datasets.
- Benchmarking: Evaluating classifiers to select the optimal model.
- External validation: Assessing generalizability across six independent external cohorts.
- Explainability: Interpreting predictive logic using SHAP-based feature attribution.
- Biological validation: Confirming biological relevance through pathway enrichment analysis.

A. Batch Effect Correction and Harmonization

To integrate multi-center datasets, the ComBat algorithm was applied to remove technical variations. PCA visualization showed that before correction, samples clustered strongly by study origin, with PC1 accounting for 57.6% of the variance [Fig. 2(a)]. After correction, these study-specific clusters merged into a unified distribution, with PC1 variance dropping to 16.3% [Fig. 2(b)]. This successful integration ensured that subsequent model training relies on shared biological signals rather than technical noise.



(a) (b)

Fig. 2. (a) PCA visualization before ComBat correction, (b) PCA visualization after ComBat correction.

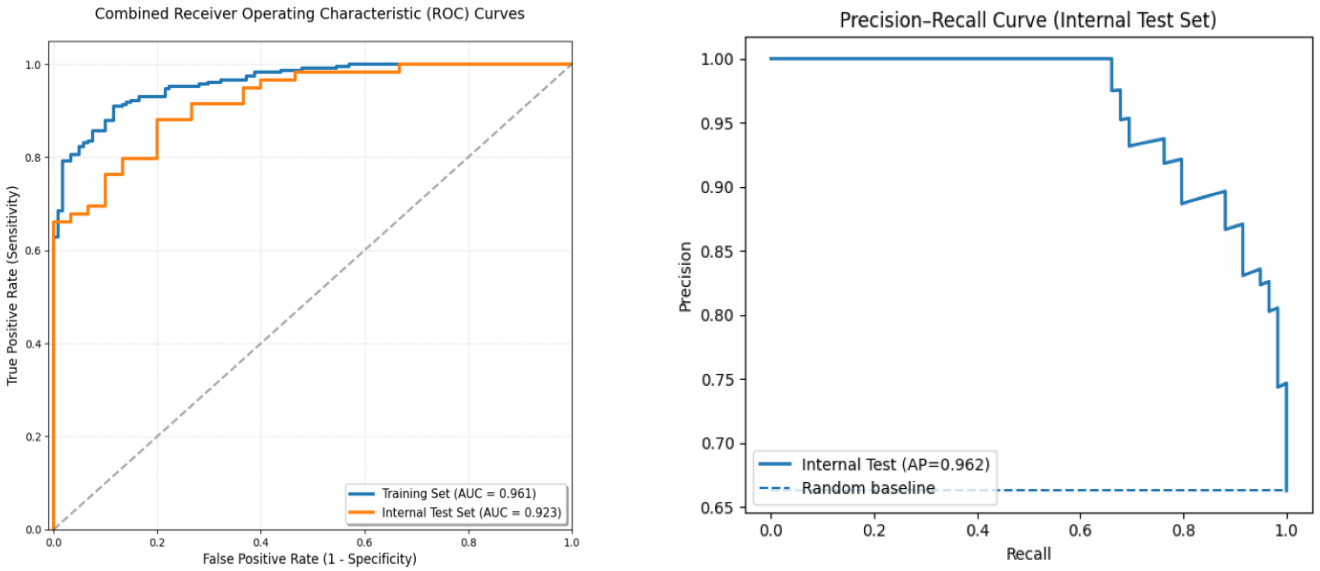
B. Predictive Performance and Benchmarking

Five classifiers were evaluated, and their performance on the internal test set is summarized in Table II.

TABLE II. INTERNAL BENCHMARKING OF CANDIDATE MODELS

Model	ROC-AUC	PR-AUC	Sens.	F1-Score	MCC	ACC
XGBoost (Selected)	**0.923**	**0.962**	**0.881**	**0.889**	**0.676**	**0.854**
Random Forest	0.954	0.977	0.847	0.901	0.749	0.876
LightGBM	0.939	0.970	0.695	0.812	0.626	0.787
Support Vector Machine	0.782	0.793	0.424	0.556	0.222	0.551
Logistic Regression	0.715	0.803	0.475	0.602	0.267	0.584

While Random Forest (RF) and LightGBM achieved marginally higher discriminative metrics (AUC of 0.954 and 0.939, respectively), XGBoost was selected as the primary engine. The model was selected for its superior sensitivity (0.881) and more stable cross-cohort generalization performance. Specifically, LightGBM required an extreme threshold (0.9999), indicating calibration instability. Despite its slightly lower AUC (0.923) [Fig. 3(a)] compared to RF, the XGBoost model demonstrated a more robust balance for minority-class detection. This was supported by a PR-AUC of 0.962 [Fig. 3(b)] and an F1-score of 0.889. Additionally, the model demonstrated a precision of 0.897, a specificity of 0.80, and a sensitivity of 0.881, yielding an MCC of 0.676. Conversely, traditional models (LR and SVM) failed to capture the complex transcriptomic signatures, yielding significantly lower performance (AUCs of 0.715 and 0.728, respectively). Notably, external validation revealed that while RF suffered from generalization collapse (e.g., zero recall in GSE60980), XGBoost maintained stability across all cohorts. This confirmed that its optimized regularization (gamma=0.3, lambda=2) effectively captured conserved biological signals instead of platform noise. Moreover, it reinforces the selection of XGBoost as the optimal tool for balancing accuracy with clinical safety [7]. Hence, XGBoost was prioritized for detailed external validation to confirm its robust clinical generalizability.



(a) (b)

Fig. 3. (a) ROC curve of the XGBoost model on the internal test dataset, (b) Precision–Recall curve for the internal test set relative to the random baseline.

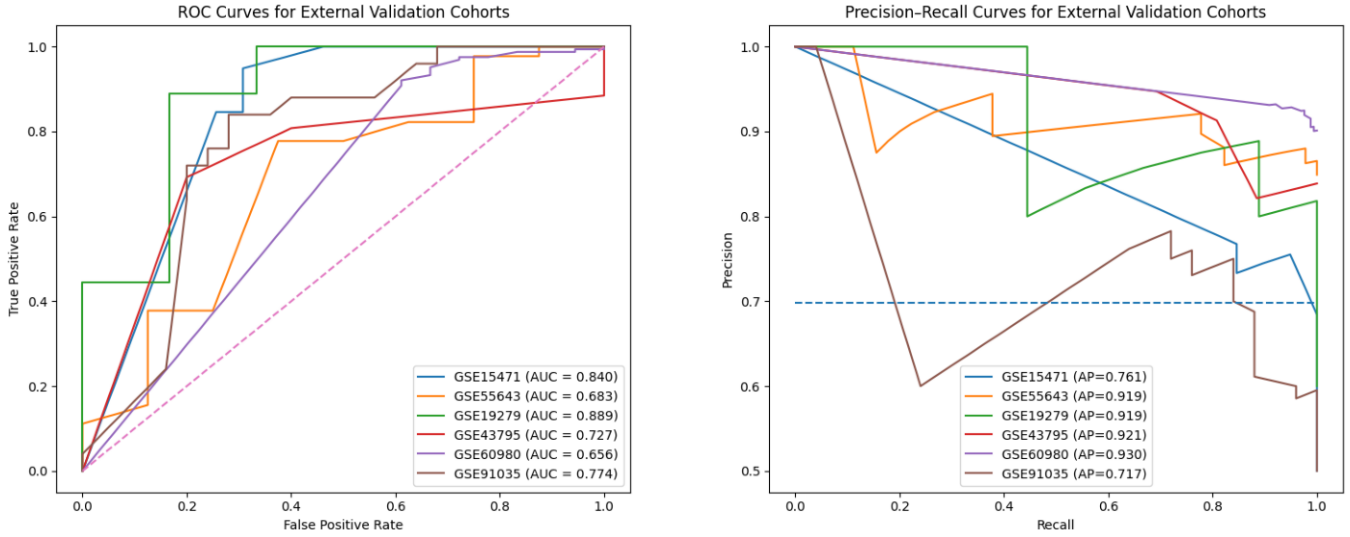
C. External Validation Across Independent Cohorts

To test the framework against real-world variability, the XGBoost model was challenged with six independent external cohorts (n=409). These datasets were kept entirely separate during the development phase to ensure an unbiased evaluation. To address platform variability, study-specific optimal thresholds were dynamically determined for each external cohort using the Youden Index. Due to platform differences, features missing in external cohorts (ranging from ~2,000 to 6,392 genes) (Table III) were excluded. The model was evaluated using only common genes. Despite this reduction, diagnostic accuracy remained high and stable across various platforms with an average AUC of 0.761.

TABLE III. COMMON GENE INTERSECTION USED FOR MODEL EXTERNAL VALIDATION

GEO Dataset	Total Genes (Original)	Common Genes with Training	Missing Genes	Gene Retention %
GSE91035	21,189	15,941	2,793	75.2%
GSE60980	22,070	16,714	2,020	75.7%
GSE43795	31,334	15,914	2,820	50.8%
GSE19279	13,515	12,342	6,392	91.3%
GSE55643	19,595	16,360	2,374	83.5%
GSE15471	23,520	18,734	4,786	79.65%

As illustrated in Fig. 4(a), the framework's performance remains consistent despite the technical noise associated with multi-center data. To provide a more comprehensive evaluation, bootstrap resampling was used (1000 iterations) to calculate 95% confidence intervals (CIs). This step confirmed the statistical robustness of the results. As shown in Table IV, the lower bounds of these CIs remain safely above the 0.5 random-chance mark for most cohorts. This demonstrated that the model is capturing real biological patterns. Also, PR-AUC, F1-score, and Matthews Correlation Coefficient (MCC) were evaluated. Precision–Recall curves are illustrated in Fig. 4(b). Overall, the model maintained strong minority-class detection across all cohorts, with PR-AUC consistently exceeding 0.71 and reaching as high as 0.93.



(a) (b)

Fig. 4. (a) ROC curves for external validation cohorts. (b) Precision–Recall curves for six external cohorts relative to the random baseline.

TABLE IV. PERFORMANCE OF THE XGBOOST MODEL ACROSS SIX INDEPENDENT COHORTS

GEO Dataset	PR-AUC	AUC [95% CI]	Acc.	Sens.	Spec.	F1-Score	MCC
GSE19279	0.919	0.889 [0.654 - 1.000]	0.867	0.889	0.833	0.889	0.722
GSE15471	0.761	0.840 [0.757 - 0.918]	0.821	0.949	0.692	0.841	0.663
GSE91035	0.717	0.774 [0.615 - 0.895]	0.780	0.778	0.783	0.792	0.564
GSE60980	0.930	0.656 [0.535 - 0.783]	0.868	0.921	0.389	0.926	0.296
GSE43795	0.921	0.727 [0.460 - 0.912]	0.710	0.692	0.800	0.800	0.372
GSE55643	0.919	0.683 [0.452 - 0.882]	0.755	0.778	0.625	0.843	0.320

Evaluation on GSE43795 and GSE55643 did not reach formal statistical significance due to the limited number of control samples (n<10). Excluding these cohorts yielded a refined average AUC of 0.790, while overall performance trends remained consistent. Overall, XGBoost captured conserved biological signals for robust cross-study prediction.

Performance variability across external cohorts was influenced by differences in cohort size, class distribution, and transcriptomic heterogeneity. Lower-performing cohorts generally contained limited control samples and reduced feature overlap after cross-platform harmonization. Despite these variations, the framework maintained stable predictive behavior across heterogeneous datasets, supporting its robustness against biological and technical variability.

In cohorts with extreme class imbalance (e.g., GSE60980 and GSE55643), MCC values were moderately reduced despite high PR-AUC and F1-scores. This reduction was mainly attributed to the limited number of normal samples rather than poor tumor detection. Nevertheless, recall remained consistently high across datasets, indicating reliable identification of cancer cases and overall framework stability.

D. SHAP-Based Model Explainability

SHAP analysis was used to interpret the XGBoost model, providing mechanistic transparency [8],[13]. As illustrated in Fig. 5, the analysis identified a subset of high-impact features. THAP9-AS1 emerged as the dominant driver, followed by RHNO1 and RUSC1-AS1. The model's reliance on long non-coding RNAs and cell-cycle-related genes aligns with

established biological dysregulation, providing a robust rationale for its high diagnostic accuracy.

E. Biological Pathway Enrichment and Mechanistic Validation

Dual-layered enrichment analysis was performed to provide a biological rationale. ORA of top SHAP features revealed a convergence on critical oncogenic hallmarks, including cell cycle progression (G2-M Checkpoint, Mitotic Spindle, and E2F Targets) and metabolic reprogramming (Glycolysis, Pentose Phosphate) (Fig. 6) [10],[11]. These findings were corroborated by GSEA, with significant enrichment in KRAS Signaling Dn (NES = -1.367; FDR = 0.045; $p < 0.001$) [Fig. 7(a)] [9],[12], Angiogenesis [Fig. 7(b)], and Unfolded Protein Response pathways [Fig. 7(c)], driving tumor adaptation to hypoxia and ER stress. Conversely, the dysregulated Xenobiotic Metabolism [Fig. 8(a)] and depletion of Pancreas Beta Cell genes [Fig. 8(b)] reflect the loss of cellular identity and chemoresistance. This concordance across independent methods confirms that the XGBoost model captures a robust, functional transcriptomic fingerprint of pancreatic malignancy rather than stochastic noise.

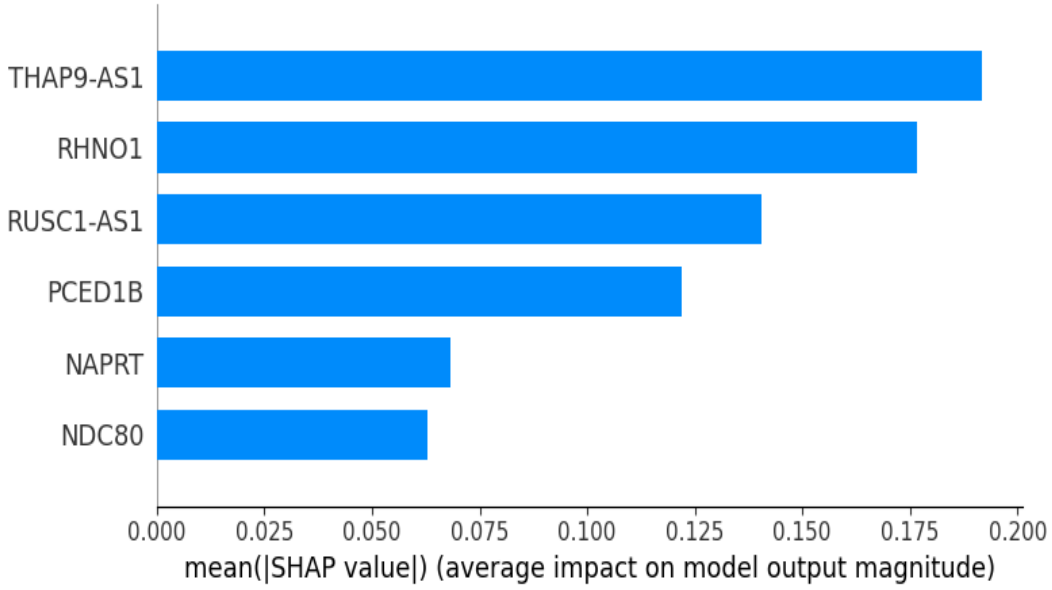


Fig. 5. Interpretability of the XGBoost model using SHAP.

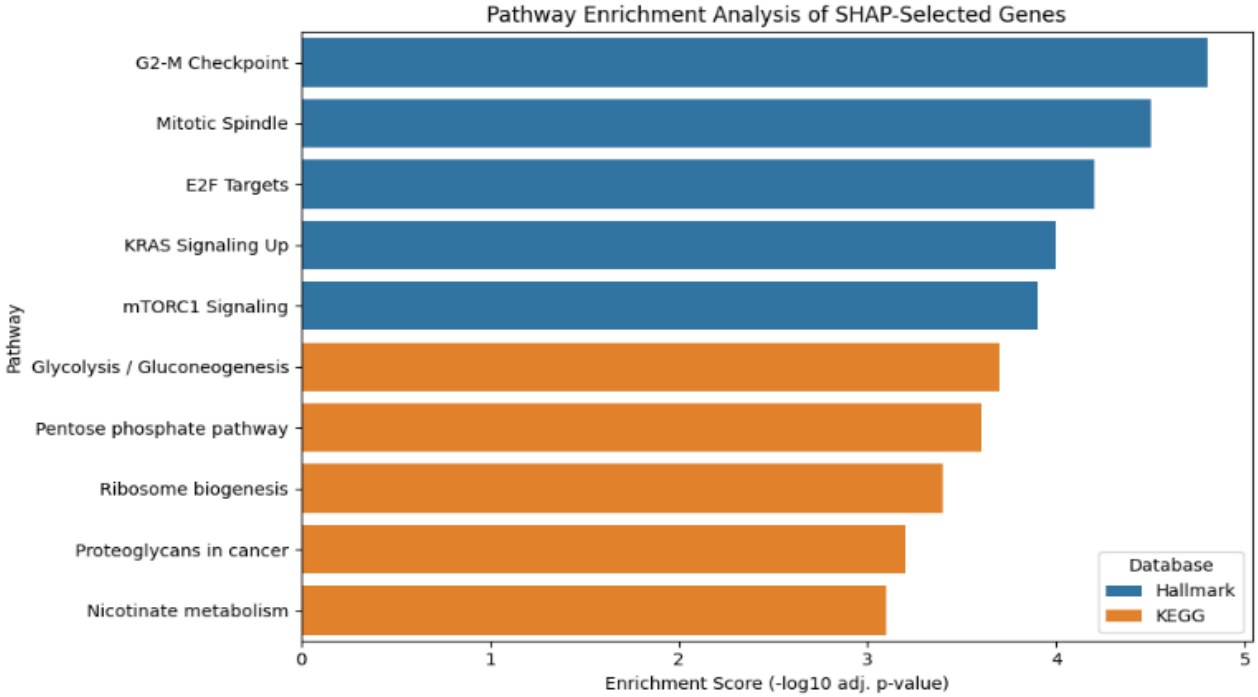
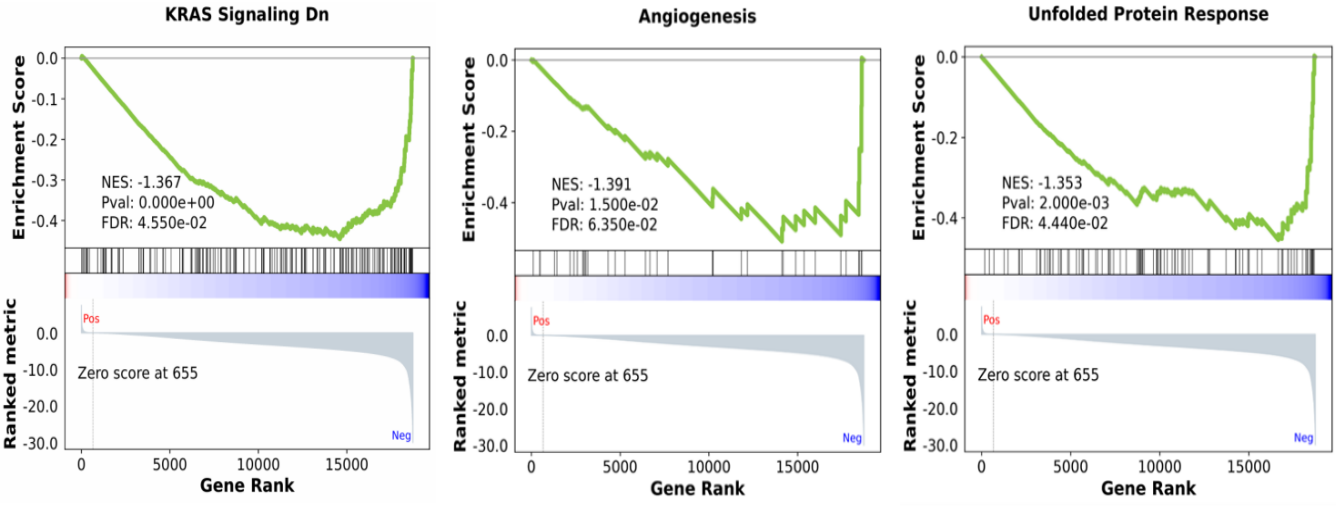
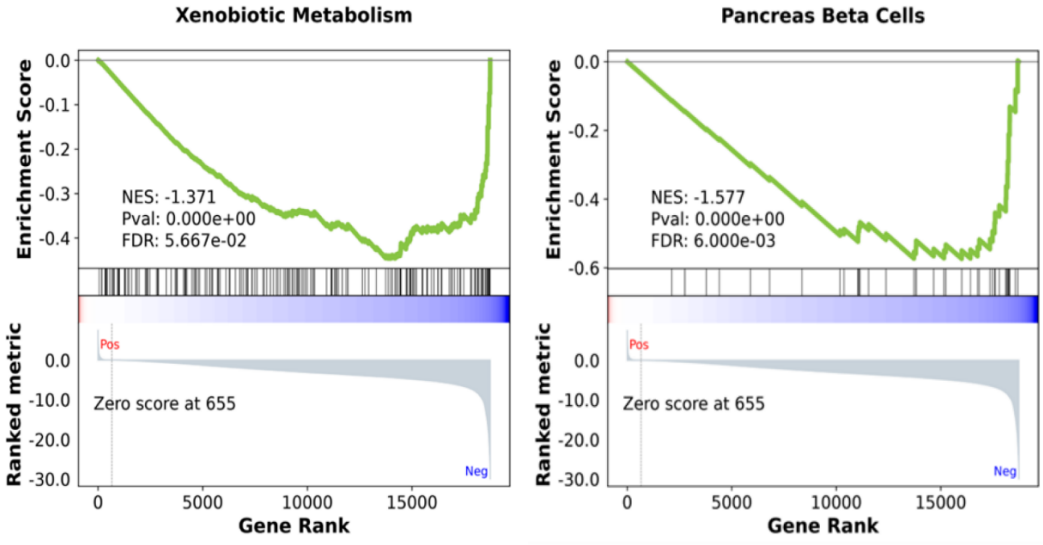


Fig. 6. ORA of SHAP-selected genes highlighting enriched Hallmark and KEGG pathways.



(a) (b) (c)

Fig. 7. (a) GSEA enrichment plot of the KRAS, (b) GSEA enrichment plot for Angiogenesis, (c) GSEA enrichment plot for UPR.



(a) (b)

Fig. 8. (a) GSEA enrichment plot for Xenobiotic metabolism, (b) GSEA enrichment plot for pancreatic beta cell signature.

VI. Discussion

XGBoost outperformed linear baselines and ensemble methods [15]. While RF and LightGBM showed higher internal AUCs, they suffered from overfitting on unseen cohorts. Given the study's emphasis on external generalizability and clinical sensitivity, final model selection was not based solely on internal discrimination performance. Internal comparison metrics alone may fail to reflect out-of-distribution overfitting across heterogeneous cohorts. Therefore, XGBoost was selected based on its superior sensitivity (0.881) and more stable predictive behavior during external validation. While class imbalance caused some variability in sensitivity, the consistently high AUC across cohorts underscores its potential for early diagnostic tools.

A key strength of this study lies in the alignment between model-driven feature importance and biological mechanisms. SHAP analysis identified influential genes [20], notably NDC80, RHNO1, and NAPRT [21]. The model also assigned importance to several long non-coding RNAs, such as THAP9-AS1 and RUSC1-AS1. However, protein-coding drivers were prioritized to facilitate a clear functional interpretation [11],[12]. NDC80, in particular, is a known regulator of cell cycle and chromosomal stability, playing a critical role in pancreatic tumor progression [21].

GSEA further validated the findings, revealing dysregulation in the cell cycle, metabolism, and KRAS signaling [12],[15],[16]. The convergence of ML feature attribution with established oncogenic mechanisms confirmed that the model captures disease-related patterns rather than technical artifacts [9],[10],[25].

Unlike in single-cohort studies, this signature remained robust across 409 unseen samples from six GEO datasets.

Prioritizing sensitivity proved effective. XGBoost maintained a stable 0.761 average AUC across all platforms. However, two cohorts (GSE43795 and GSE55643) contain fewer than ten control samples, limiting their statistical power. To prevent inflating the generalizability claims, the model was re-evaluated by excluding these two small datasets. The remaining four high-powered cohorts (325 samples) yielded a refined average AUC of 0.790. This stable performance confirmed that the model captured a fundamental biological pattern rather than sample-size artifacts or platform-specific noise.

Despite these results, limitations exist. While ComBat reduced batch effects [3],[4], residual variability may persist due to platform-specific differences in probe sensitivity. Nevertheless, ComBat remains the gold standard for multi-platform data harmonization. Recent benchmarks demonstrate its superior ability to preserve biological signals compared to complex deep-learning alternatives [24]. Additionally, reliance on public datasets limits performance assessment in prospective clinical settings.

The study lacks direct experimental validation; findings for NDC80 and THAP9-AS1 remain entirely in silico. To guide future wet-lab translation, a concrete three-phase workflow is proposed.

First, baseline expression levels will be quantified via RT-qPCR across pancreatic cancer cell lines, such as PANC-1 and MIA PaCa-2, versus normal HPDE controls. Western blotting will further confirm NDC80 protein translation. Second, functional loss-of-function assays will evaluate phenotypic changes using siRNA or shRNA-mediated knockdown. Specifically, MTT/CCK-8 proliferation and Transwell invasion assays will determine if depleting these targets suppresses oncogenic capabilities. Finally, ex vivo validation using immunohistochemistry (IHC) on patient tissue microarrays will map NDC80 expression across histological tumor grades. Mechanistically, future assays should clarify whether the lncRNA THAP9-AS1 acts as a competitive endogenous RNA (ceRNA) to regulate NDC80.

Furthermore, integrating clinical covariates like CA 19-9 could mitigate specificity fluctuations in imbalanced cohorts [14],[17]. Future work incorporating longitudinal data and multimodal biomarkers is essential to enhance predictive robustness and realize the framework's full translational potential [22].

VII. Conclusion

In this study, an interpretable machine learning framework was presented for pancreatic cancer prediction. Systematic batch-effect correction and comparative benchmarking of five classifiers identified XGBoost as the most clinically robust model. The proposed framework achieved strong predictive performance and maintained generalizability across multiple external cohorts. SHAP-based explainability enabled the identification of key molecular features associated with pancreatic tumorigenesis. These findings were further supported through pathway enrichment analyses, highlighting established oncogenic mechanisms including cell cycle dysregulation and KRAS signaling.

Overall, the findings demonstrated the potential of explainable machine learning for extracting biologically grounded signatures from high-dimensional transcriptomic data. However, the proposed framework should currently be considered an exploratory computational foundation. Prospective clinical validation and further clinical integration studies remain necessary before translational application in real-world diagnostic settings.

Data and Code Availability

The datasets analyzed in this study are publicly available from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). The code used in this study is available from the corresponding author upon reasonable request.

References

[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.

[2] Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A., “Cancer statistics, 2023,” CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17-48, 2023.

[3] J. T. Leek et al., “The sva package for removing batch effects and other unwanted variation in high-throughput experiments,” Bioinformatics, vol. 28, no. 6, pp. 882–883, 2012.

[4] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” Biostatistics, vol. 8, no. 1, pp. 118–127, 2007.

[5] F. A. Almisned, N. Usanase, D. U. Ozsahin, and I. Ozsahin, "Incorporation of explainable artificial intelligence in ensemble machine learning-driven pancreatic cancer diagnosis," *Scientific Reports*, vol. 15, no. 1, p. 14038, 2025, doi: 10.1038/s41598-025-98298-0.

[6] F. Carrillo-Perez et al., "Machine learning identifies 6-gene signature in peripheral blood for pancreatic cancer diagnosis," Heliyon, vol. 11, no. 1, p. e43138, Jan. 2025, doi: 10.1016/j.heliyon.2025.e43138.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf., pp. 785–794, 2016.

[8] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[9] A. Subramanian et al., “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” Proc. Natl. Acad. Sci. USA, vol. 102, no. 43, pp. 15545–15550, 2005.

[10] M. Ashburner et al., “Gene ontology: Tool for the unification of biology,” Nature Genetics, vol. 25, no. 1, pp. 25–29, 2000.

[11] Cancer Genome Atlas Research Network, “Integrated genomic characterization of pancreatic ductal adenocarcinoma,” Cancer Cell, vol. 32, no. 2, pp. 185–203, 2017.

[12] P. Bailey et al., “Genomic analyses identify molecular subtypes of pancreatic cancer,” Nature, vol. 531, no. 7592, pp. 47–52, 2016.

[13] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.

[14] L. Rahib et al., “Projected incidence and deaths of pancreatic cancer in the United States, 2010–2030,” Cancer Research, vol. 74, no. 11, pp. 2913–2921, 2014.

[15] A. M. Waters and C. J. Der, “KRAS: The critical driver and therapeutic target in pancreatic cancer,” Cold Spring Harbor Perspectives in Medicine, vol. 8, no. 9, a031435, 2018.

[16] J. Gu et al., "Selection of key biomarkers and functions of pancreatic adenocarcinoma based on multiple internal and external data sets," Frontiers in Genetics, vol. 12, p. 745370, 2022.

[17] L. Chen et al., "A systematic evaluation of methods for correcting batch effects in transcriptomic data," Briefings in Bioinformatics, vol. 22, no. 4, bbaa348, 2021.

[18] X. Chen et al., "Identification of a cell cycle-related gene signature for predicting survival in pancreatic cancer," Cancer Medicine, vol. 11, no. 8, pp. 1914-1926, 2022.

[19] G. J. Rao and A. S. Prasad, "Classification of gene expression from RNA-seq data for pancreatic cancer prognosis using ensemble learning," Journal of Applied Biology & Biotechnology, vol. 12, no. 3, pp. 45–53, May-Jun. 2024, doi: 10.7324/JABB.2024.171755.

[20] K. S. Babu et al., "SHAP-based explanation of a machine learning model for predicting pancreatic cancer," Informatics in Medicine Unlocked, vol. 32, p. 101026, 2022.

[21] T. Liu et al., "NDC80 promotes proliferation and metastasis of pancreatic cancer through the Wnt/β-catenin signaling pathway," Journal of Cancer, vol. 12, no. 16, pp. 4843–4853, 2021.

[22] A. K. Shukla et al., "Integrative analysis of multi-omics data for biomarker discovery in cancer," Briefings in Bioinformatics, vol. 22, no. 4, bbaa212, 2021.

[23] J. Ge, J. Cai, G. Zhang, D. Li, and L. Tao, "Multi-omics integration and machine learning uncover molecular basal-like subtype of pancreatic cancer and implicate A2ML1 in promoting tumor epithelial-mesenchymal transition," Journal of Translational Medicine, vol. 23, no. 1, p. 741, 2025.

[24] Yu, Y., Zhang, N., Mai, Y. et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. Genome Biol 24, 201 (2023).

[25] The Gene Ontology Consortium, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, 2023, doi: 10.1093/genetics/iyad031.

A Hybrid PCA–Random Forest Model for Predicting Employee Performance from Work Ethics and Work Values

Nova E. Arquillano
Department of Public Administration, Pangasinan State University, Philippines

***Abstract*—This study examined the predictive relationship of work ethics and work values with employee task performance using a hybrid PCA–Random Forest model. Data were obtained from 231 PSU employee respondents and 81 questionnaire variables covering task performance, work ethics, and work values. A quantitative, cross-sectional predictive design was used; therefore, the findings are interpreted as predictive associations rather than causal effects. Descriptive statistics, reliability analysis, correlation analysis, Principal Component Analysis (PCA), baseline regression models, Random Forest regression, and a hybrid PCA–Random Forest model were applied. PCA reduced the 74 predictor items into 19 components, explaining 80.9% of the predictor variance. On the held-out test set, the hybrid PCA–Random Forest model achieved MAE = 0.272, RMSE = 0.383, and R^2 = 0.524. Standard Random Forest produced a similar test performance (MAE = 0.283, RMSE = 0.383, R^2 = 0.526), indicating that PCA did not substantially improve accuracy but produced a more compact and less redundant feature representation. Professionalism, commitment to public interest, and nationalism and patriotism emerged as important predictors of task performance. The study demonstrates the usefulness of interpretable machine-learning models for evidence-based human resource development in public higher education, while noting limitations related to ceiling effects, self-report data, single-institution sampling, and the need for external validation.**

***Keywords*—Employee performance; work ethics; work values; principal component analysis; random forest; hybrid model; machine learning; PSU employees**

I. Introduction

Employee performance is a critical concern in public sector institutions because the quality of public service depends largely on how effectively employees perform their duties. In state universities, employee performance is commonly reflected in task completion, timeliness, efficiency, and quality of outputs, all of which directly affect institutional effectiveness and stakeholder satisfaction. In the present study, employee performance is operationalized through task performance indicators such as planning work on time, accomplishing tasks efficiently, and producing quality outcomes, while work ethics and work values are treated as key explanatory factors.

Work ethics and work values are particularly important in the context of public employment because they shape employees' sense of accountability, professionalism, integrity, responsiveness, and commitment to the public interest. Ethical conduct and values-based behavior influence not only how employees carry out assigned responsibilities but also how they interact with clients, colleagues, and the organization as a whole. Prior studies have shown that workplace attitudes and behavioral factors are significantly related to task performance and work outcomes [1], while innovative work behavior and citizenship-related behavior also contribute to stronger workplace contribution and effectiveness [2]. In addition, job satisfaction and work engagement have been linked to desirable organizational outcomes and reduced turnover intention, suggesting that employee values and work orientation play a meaningful role in performance-related behavior [3].

Despite the recognized importance of work ethics and work values, modeling their relationship with employee performance remains methodologically challenging. These constructs are usually measured through multiple Likert-scale items, making them inherently multidimensional and often highly correlated. Traditional statistical methods remain useful for explanation, but they may be limited in capturing nonlinear relationships and interaction effects among behavioral variables. Recent work has emphasized that machine learning models are increasingly suitable for the social sciences because they can handle complex and nonlinear associations more effectively than strictly linear techniques [4], [5]. Best-practice discussions in psychological and behavioral machine learning further note that supervised learning methods can strengthen prediction when data preparation, validation, and interpretation are conducted rigorously [6].

In psychometric and survey-based studies, the quality of indicators strongly affects model performance and interpretability. Multivariate measurement research has long emphasized that the selection and treatment of indicators can substantially influence modeling outcomes [7]. This is particularly relevant in employee studies where ethics, values, and performance are each represented by multiple observed variables rather than by a single direct measure. Psychometric evaluation studies also demonstrate the importance of sound measurement properties in survey-based research [8], [9]. For this reason, any predictive study involving questionnaire data should pay careful attention to the structure, redundancy, and quality of indicators before model construction.

Another major concern in survey-based modeling is data quality. Missing responses, careless answering, and meaningless response patterns may reduce the reliability of results if left untreated. Missing data have been identified as a key issue in survey research, with consequences for estimation quality, validity, and interpretation [10]. Likewise, nonreactive

indicators can help identify poor-quality or meaningless responses in self-report surveys [11]. Because the present study relies on multiple questionnaire items measuring work ethics, work values, and task performance, careful preprocessing is necessary to ensure that the predictive model is based on valid and stable information.

Given the large number of survey indicators, dimensionality reduction becomes a practical necessity. Principal Component Analysis (PCA) is a widely used technique for transforming correlated variables into a smaller number of components while preserving as much information as possible. This approach is especially useful when many predictors overlap conceptually and statistically, as is common in behavioral and organizational datasets [7], [12]. By reducing redundancy and multicollinearity, PCA can improve the efficiency of downstream predictive models and simplify interpretation at the component level. In multivariate research, such reduction techniques are valuable when latent structures must be approximated from many observed indicators [13].

After dimensionality reduction, Random Forest is an appropriate predictive algorithm for employee performance because it can accommodate nonlinear relationships, higher-order interactions, and complex variable patterns without strict distributional assumptions. Machine learning research in psychometrics and the social sciences has shown that such methods are particularly advantageous when the goal is prediction rather than only parameter estimation [4], [5], [6]. Random Forest is also robust for structured tabular data and can perform well in moderate-sized datasets such as employee survey data. Since work ethics and work values may influence performance in nonuniform ways across respondents, an ensemble-based learner such as Random Forest is a suitable modeling choice.

A further strength of Random Forest is its compatibility with model interpretation tools. In applied organizational research, predictive accuracy alone is insufficient; decision makers also need to understand which factors are driving model predictions. Interpretable machine learning frameworks provide ways to explain black-box models and support evidence-based decision-making [14]. Feature attribution methods such as Shapley-based explanations can be used to estimate the contribution of individual predictors or components to model output [15], [16]. This is especially valuable in the present study because it allows the identification of which ethical and value-based dimensions are most strongly associated with employee performance.

Model evaluation is also essential in determining whether a predictive framework is appropriate and defensible. Comparative assessment using established evaluation metrics is recommended in machine learning studies to ensure that predictive performance is examined systematically and transparently [17]. In social science applications, methodological rigor likewise requires proper training-validation procedures and responsible reporting of predictive results [6]. Consequently, the adoption of a hybrid model should not only be justified conceptually but also supported empirically through relevant performance measures.

The use of a hybrid PCA–Random Forest framework is therefore justified both statistically and substantively. Statistically, PCA addresses the high dimensionality and multicollinearity common in questionnaire-based predictors, while Random Forest captures nonlinear patterns that may exist between employee ethics, values, and task performance [4], [7], [12]. Substantively, the model supports a more data-driven understanding of how employee characteristics translate into work outcomes in a public university setting. Related studies have shown that work-related attitudes, engagement, and behavioral orientations are connected to performance, satisfaction, and organizational contribution [3], [18], [19]. These findings reinforce the importance of examining performance not only as a function of technical capacity but also as an outcome associated with ethical and value-based dimensions of work.

The present study is also timely because organizational environments are becoming increasingly data-driven, requiring more robust methods for analyzing human-centered workplace data. Machine learning has gained growing relevance in social and behavioral inquiry because it can complement conventional methods in identifying predictive structures from complex datasets [4]. Tutorials and methodological reviews likewise encourage the careful adoption of supervised machine learning in psychology and related disciplines to enhance predictive validity and methodological innovation [17]. In this regard, the proposed hybrid model contributes to the emerging application of machine learning in educational and public sector workforce studies.

Accordingly, this study proposes A Hybrid PCA–Random Forest Model for Predicting Employee Performance from Work Ethics and Work Values among Employees. Specifically, it seeks to determine whether the multidimensional indicators of work ethics and work values can be reduced into meaningful components and used to predict employee task performance with acceptable accuracy and interpretability. By integrating survey-based measurement, dimensionality reduction, ensemble learning, and interpretable machine learning, the study contributes both a methodological approach and an applied perspective for understanding employee performance [14].

Although previous studies have established associations between work-related attitudes, values, engagement, and performance, much of the existing literature remains descriptive or correlational. Fewer studies have examined whether multidimensional work ethics and work values indicators can predict employee performance using a transparent machine-learning framework. In addition, many predictive studies emphasize accuracy but provide limited interpretation of the psychological or organizational meaning of model features. This creates a gap for an approach that combines dimensionality reduction, nonlinear prediction, and interpretability. The present study addresses this gap by evaluating whether PCA-derived dimensions of work ethics and work values can be used in a Random Forest model to predict task performance among public university employees. Ultimately, the findings may provide evidence-based support for institutional strategies aimed at strengthening employee effectiveness through ethics- and values-oriented development.

II. METHODOLOGY

This section presents the research design, data sources, and analytical procedures used in the study. It describes how the predictor variables were measured and how the six regression models were developed and compared to determine the most appropriate approach for analyzing the relationships among the variables.

A. Research Design

This study employed a quantitative, cross-sectional, predictive research design. The objective was to develop and evaluate a hybrid PCA–Random Forest model for predicting employee performance from work ethics and work values among Pangasinan State University (PSU) employees. The predictive framework was selected because machine learning methods are well suited for social science data with many correlated variables and potentially nonlinear relationships. The hybrid design combined Principal Component Analysis (PCA) for feature reduction with Random Forest Regression for prediction, a combination that is appropriate when survey-based predictors are numerous, overlapping, and multidimensional. The study was explanatory-predictive in nature. It not only describes the levels of work ethics, work values, and employee performance, but also models how ethics- and value-related indicators jointly predict task performance. This design is consistent with current recommendations that machine learning in behavioral and psychometric research should be used to complement traditional explanatory approaches through rigorous preprocessing, validation, and interpretable modeling. The study is therefore best interpreted as a predictive and exploratory model of performance-related patterns within the PSU employee sample.

B. Data Source and Unit Analysis

The dataset used was taken from the survey of PSU employees, which contains 231 respondent records and 81 item-level variables corresponding to the questionnaire dimensions. The unit of analysis was the individual PSU employee respondent. The predictor set consisted of work ethics and work values items, while the outcome variable was employee task performance derived from the task productivity indicators in the questionnaire. The questionnaire is composed of three major parts. Part I measures task performance through seven indicators (pro1–pro7) using a 5-point agreement scale. Part II measures work ethics using multiple indicators across eight dimensions: commitment to public interest, professionalism, political neutrality, justness and sincerity, responsiveness to the public, nationalism and patriotism, commitment to democracy, and simple living. Part III measures work values across six dimensions: God-fearing, accountability, integrity, transparency, innovativeness, and competency. These constructs and response formats are explicitly defined in the uploaded instrument.

Because the data were collected through self-report questionnaires, the findings may be affected by common-method bias and socially desirable responding. This concern is particularly relevant because ethics and values are socially valued constructs. The results should therefore be interpreted as respondents' reported perceptions of their work behavior and values rather than as direct administrative performance records.

C. Research Locale and Population

The study focused on PSU employees as reflected in the survey instrument. The target population, therefore, consisted of PSU personnel who responded to the questionnaire. The study treated the dataset as the empirical basis for training and evaluating the predictive model.

Because the data were collected through self-report questionnaires, the findings may be affected by common-method bias and socially desirable responding. This concern is particularly relevant because ethics and values are socially valued constructs. The results should therefore be interpreted as respondents' reported perceptions of their work behavior and values rather than as direct administrative performance records.

D. Instrumentation and Variable Operationalization

Employee performance was operationalized as task performance, based on the seven indicators in Part I of the questionnaire: planning work on time, optimal planning, distinguishing major from minor issues, performing work with minimal time and effort, completing tasks efficiently, producing high-quality outputs, and keeping intended results in mind. These seven indicators were rated on a 5-point Likert-type scale and were combined to form the target construct for prediction.

The predictor variables were drawn from Parts II and III of the questionnaires. The work ethics construct includes the following dimensions and item groups: commitment to public interest (com1–com7), professionalism (prof1–prof7), political neutrality (po1–po4), justness and sincerity (j1–j5), responsiveness to the public (res1–res6), nationalism and patriotism (na1–na6), commitment to democracy (co1–co3), and simple living (sim1–sim6). The work values construct includes God-fearing (god1–god5), accountability (acc1–acc5), integrity (in1–in5), transparency (tra1–tra5), innovativeness (inno1–inno5), and competency (compe1–compe5, including the dataset's comp3 field corresponding to the third competency item). These domains and items are specified in the questionnaire. All predictors were treated as ordinal Likert-type responses coded numerically from 1 to 5. In practice, Likert item responses are commonly modeled as numeric inputs in psychometric and supervised learning workflows, particularly when the goal is prediction rather than purely ordinal inference.

E. Data Preparation and Screening

Fig. 1 shows the data preprocessing and screening flowchart. The dataset was screened for completeness, correct variable names, duplicate cases, and valid response ranges. Because all questionnaire items used a 1–5 scale, any value outside this range was treated as invalid and corrected or removed when necessary. Missing data were examined at both the item and respondent levels. Cases with excessive missing responses were excluded, while minimal random missingness was handled through item-median imputation. Composite means were computed only when most items in a dimension were available. The data were also checked for low-quality response patterns, such as straight-lining, invariant responses, and duplicated answer sets, and clearly invalid cases were removed. Finally, all predictor variables were standardized into z-scores before PCA to ensure comparability and prevent variables with larger variance from dominating the component extraction.

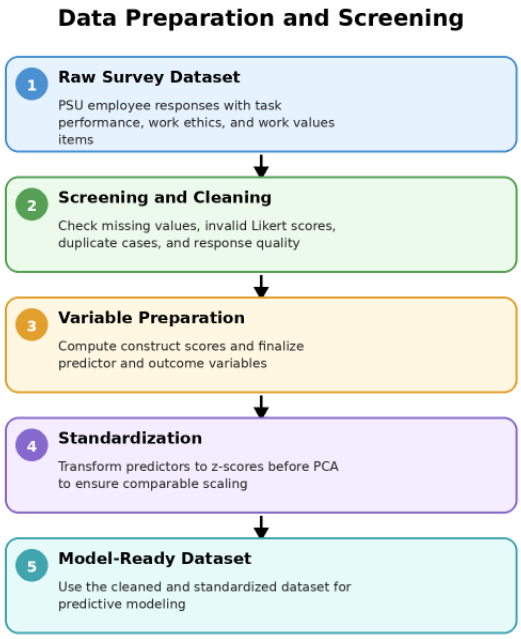


Fig. 1. Data preparation and screening.

F. Feature Engineering

Two parallel feature representations were used in the study: item-level predictors: all work ethics and work values items entered individually; and PCA-derived predictors: principal components extracted from the full predictor matrix.

The principal analysis used the PCA-derived feature space because the study proposes a hybrid PCA–Random Forest framework. The item-level representation was retained as a benchmark to determine whether PCA improved, reduced, or preserved predictive performance compared with a standard Random Forest using the original survey items.

G. Principal Component Analysis (PCA)

PCA was applied to the standardized predictor matrix composed of all work ethics and work values items. The purpose of PCA was to reduce redundancy among correlated survey items and derive a smaller number of orthogonal components that summarize the underlying variation in employees' ethics- and values-related responses. PCA is appropriate in multivariate measurement settings where many observed indicators capture overlapping constructs. Using multiple retention criteria is preferable because purely mechanical cutoffs may either retain noisy components or discard useful information. The retained principal components then served as the input features for the Random Forest model. This PCA stage helps reduce multicollinearity, compress noise, and improve model stability when many survey items are conceptually related.

The use of PCA before Random Forest was not based on the assumption that Random Forest requires uncorrelated predictors. Rather, PCA was used to address the high redundancy of the questionnaire indicators and to create a compact representation of the work ethics and work values domains. Since Random Forest can already handle correlated inputs, the hybrid model was evaluated against standard Random Forest to determine whether dimensionality reduction added practical value. This comparison is important because PCA may also introduce information loss when components do not preserve predictive patterns in the original variables.

H. Random Forest Regression

The predictive model used in the second stage was Random Forest Regression. Random Forest is an ensemble learning algorithm that builds multiple decision trees from bootstrap samples and aggregates their outputs to produce a final prediction. It is especially suitable for structured social science datasets because it can model nonlinear effects and variable interactions without requiring strict linearity or normality assumptions. The choice of Random Forest was justified because it handles moderate sample sizes effectively, it is robust to overfitting compared with single decision trees, it can capture nonlinear relationships between ethics, values, and performance, and it supports feature importance and post hoc interpretability analysis. Fig. 2 shows the model implementation process.

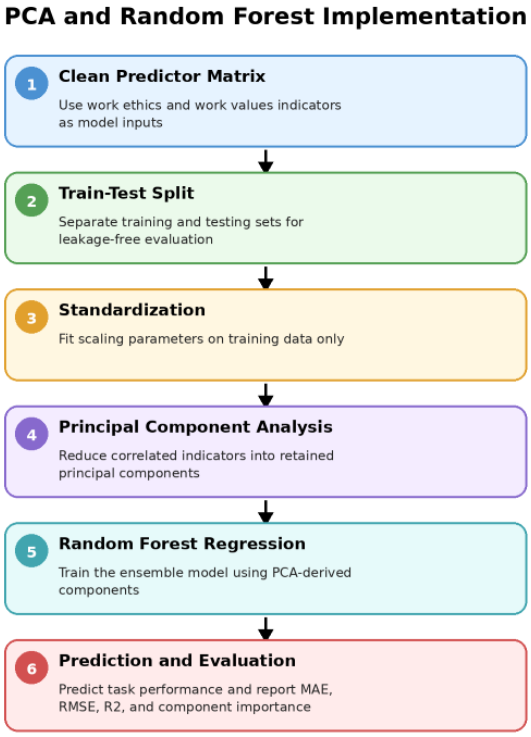


Fig. 2. PCA and random forest implementation.

I. Train–Test Procedure

To evaluate generalization performance, the dataset was partitioned into training and testing subsets. A practical split for the present sample is 80% training and 20% testing, with the random seed fixed for reproducibility. Model development, hyperparameter tuning, and PCA fitting were performed only on the training set to prevent information leakage. The trained transformation and model were then applied to the held-out test set. More specifically, the pipeline followed these steps:

- split the full dataset into training and testing sets,
- fit the standardization parameters using the training set only,

- fit PCA using the standardized training predictors only,
- transform both training and testing predictors using the fitted PCA model,
- train the Random Forest regressor on the transformed training set,
- generate predictions on the transformed test set.

This leakage-free workflow is aligned with best practices in supervised machine learning.

J. Hyperparameter Tuning

Random Forest performance depends on several tuning parameters. The following hyperparameters were optimized through k-fold cross-validation on the training set:

- number of trees (n_estimators),
- maximum tree depth (max_depth),
- minimum samples required to split a node (min_samples_split),
- minimum samples required at a leaf (min_samples_leaf),
- number of predictors considered at each split (max_features).

A 5-fold cross-validation procedure is suitable for the present sample size because it balances bias and variance in performance estimation. Grid search or randomized search may be used to identify the best-performing parameter combination. This is consistent with methodological recommendations for machine learning applications in psychology and social research.

K. Baseline and Comparative Models

To demonstrate the value of the hybrid approach, the PCA–Random Forest model should be compared against baseline models. The recommended comparison set includes:

- multiple linear regression,
- ridge regression or elastic net,
- decision tree regression,
- random forest without PCA.

Linear and regularized regression models provide interpretable baselines and are commonly recommended when working with many correlated predictors. Comparing the hybrid model against these alternatives allows the study to show whether PCA improves the predictive performance of Random Forest and whether nonlinear ensemble learning performs better than simpler linear approaches.

L. Model Evaluation Metrics

Model performance was evaluated using standard regression metrics. The principal metrics were Mean Absolute Error, Root Mean Squared Error, and Coefficient of Determination.

Robustness was examined by comparing cross-validation results with held-out test performance and by comparing the hybrid PCA–Random Forest model with standard Random Forest. Similar performance across cross-validation and test results was treated as evidence that the model was not merely fitting noise in one split of the data. Because the sample size was moderate, model complexity was controlled through cross-validation and hyperparameter tuning.

M. Model Interpretation

Because Random Forest is an ensemble model, interpretability must be explicitly addressed. The study, therefore, included a model explanation phase after prediction. Two levels of interpretation were proposed:

- Global interpretation, using permutation importance or impurity-based feature importance to determine which retained PCA components contributed most strongly to prediction,
- Local interpretation, using SHAP-based or Shapley-style explanation methods to estimate how features influenced individual predictions.

Interpretable machine learning is important because predictive accuracy alone is insufficient in applied organizational settings. Administrators need to know which dimensions of work ethics and work values are most strongly associated with employee performance in order to translate findings into policy or intervention. When PCA components are used, the interpretation should proceed in two stages: first, identify important components, then inspect the original variable loadings within those components to understand the substantive meaning of the prediction drivers.

For the hybrid model, interpretation was conducted by examining the importance of each retained principal component and then reviewing the original item loadings within the most influential components. Components were labelled based on the dominant work ethics and work values indicators that loaded on them. This procedure allowed the statistical output of PCA to be translated back into substantive organizational constructs.

N. Reliability and Measurement Assessment

Although the main contribution of the study is predictive, reliability and measurement adequacy should still be examined. For each major questionnaire dimension, Cronbach's alpha or an equivalent internal consistency statistic may be computed. This step helps establish that the grouped indicators used in the analysis show acceptable coherence. Psychometric evaluation remains important even in machine learning workflows because poor measurement quality can undermine predictive validity. In addition, descriptive statistics for each construct should be reported, including mean, standard deviation, minimum, maximum, skewness, and kurtosis. A correlation matrix may also be presented to show the extent of interdependence among the item groups before PCA.

O. Software Implementation

The analysis may be implemented in Python using standard scientific computing libraries:

- pandas for data handling,
- numpy for numerical operations,

- scikit-learn for standardization, PCA, train–test split, cross-validation, and Random Forest,
- matplotlib or seaborn for visual outputs,
- shap for model interpretation.

Python is appropriate because it supports end-to-end preprocessing, hybrid modeling, evaluation, and explainability within one reproducible workflow.

P. Ethical Considerations

The study used questionnaire responses from PSU employees and, therefore, involved human-subject data. Ethical handling of the dataset required that all records be analyzed in aggregate or anonymized form, with no direct personal identifiers retained in the modeling file. Access to the raw data should be limited to the researcher, and findings should be reported only at the level of patterns, not individual identities. These safeguards are especially important in workplace studies involving ethics, values, and performance.

III. Results and Discussions

This section presents the findings of the study and provides an interpretation of the results based on the data analysis. It highlights the patterns observed in the predictor constructs and compares the performance of the different regression models to explain their implications in relation to the objectives of the study.

A. Data Screening and Preliminary Analysis

The dataset contained 231 complete responses and 81 variables. No missing values were detected, and all item responses were within the valid Likert range of 1 to 5. This indicates that the dataset was already clean enough for predictive modeling without requiring imputation or deletion. A notable pattern in the data was the strong concentration of responses at the upper end of the scale. Across all item responses, about 69.8% were rated 5, 24.8% were rated 4, and only about 5.4% fell in the 1–3 range. This suggests that PSU employees generally rated themselves highly on work ethics, work values, and performance. At the same time, this pattern points to a likely ceiling effect, meaning the predictive models had to distinguish performance using relatively compressed high-score data rather than widely dispersed responses. The composite task performance score had a mean of 4.37 and a standard deviation of 0.50, with observed scores ranging from 2.86 to 5.00. This indicates that employee performance was generally high but still variable enough to support prediction.

The ceiling effect is an important limitation of the dataset. Since most responses were concentrated at the upper end of the scale, the models had less variation available for distinguishing respondents with different performance levels. This may reduce discrimination capability and may also explain why some normatively important constructs, such as integrity and transparency, were not among the strongest statistical predictors.

B. Reliability of the Scales

Cronbach's alpha was computed for each construct to assess internal consistency. All scales showed acceptable to excellent reliability. Table I shows the summary of the reliability test results.

TABLE I. Test of Reliability

Construct	Items	Mean	SD	Cronbach's alpha
Task performance	7	4.372	0.501	0.856
Commitment to public interest	7	4.417	0.506	0.743
Professionalism	7	4.552	0.428	0.843
Political neutrality	4	4.65	0.566	0.775
Justness and sincerity	5	4.764	0.413	0.901
Responsiveness to the public	6	4.689	0.399	0.862
Nationalism and patriotism	6	4.542	0.448	0.829
Commitment to democracy	3	4.641	0.531	0.922
Simple living	6	4.478	0.555	0.842
God-fearing	5	4.732	0.579	0.866
Accountability	5	4.718	0.439	0.9
Integrity	5	4.842	0.278	0.766
Transparency	5	4.808	0.347	0.852
Innovativeness	5	4.737	0.374	0.832
Competency	5	4.781	0.346	0.832

Table I reports reliability results for all major task performance, work ethics, and work values dimensions, including the number of items, mean, standard deviation, and Cronbach's alpha for each scale. This complete subscale reporting was added to allow readers to judge the internal consistency and measurement quality of every dimension included in the predictive model.

These findings show that the questionnaire dimensions were sufficiently coherent for both descriptive analysis and machine-learning modeling. The strongest reliability estimates were found for commitment to democracy ($\alpha = 0.922$), justness and sincerity ($\alpha = 0.901$), and accountability ($\alpha = 0.900$), while the lowest, though still acceptable, was commitment to public interest ($\alpha = 0.743$).

C. Descriptive Results

Among the predictor constructs, the highest mean scores were recorded for integrity (4.842), followed by transparency (4.808), competency (4.781), justness and sincerity (4.764), and innovativeness (4.737). In contrast, the lowest mean scores, although still relatively high, were found in task performance (4.372), commitment to public interest (4.417), simple living (4.478), nationalism and patriotism (4.542), and professionalism (4.552). This pattern suggests that PSU employees rated themselves most strongly in terms of personal and ethical standards, particularly integrity, transparency, and competency, while somewhat lower ratings were observed in more behaviorally demanding or institutionally oriented dimensions such as task performance and commitment to public interest. Substantively, this may indicate that respondents viewed themselves as highly principled and capable, but less uniformly

strong in consistently translating these values and orientations into high task-performance outcomes. Since all mean scores are well above 4.0, the issue does not appear to be weak ethics or poor performance, but rather which dimensions are most effective in explaining differences among an already high-performing group of respondents.

D. Bivariate Relationship with Task Performance

Table II shows the strongest positive correlations with task performance.

TABLE II. SUMMARY OF CORRELATION

Predictor construct	Correlation with task performance
Professionalism	0.455
Commitment to public interest	0.418
Nationalism and patriotism	0.36
Justness and sincerity	0.346
Responsiveness to the public	0.334
Accountability	0.314
Political neutrality	0.313
Commitment to democracy	0.305
Simple living	0.265
Integrity	0.241

These results suggest that professionalism had the strongest linear association with employee performance, followed by commitment to public interest. This is intuitive because the task-performance indicators focus on planning, efficiency, quality, and work execution, which are behaviorally close to professional conduct and duty orientation.

The weaker, though still positive, association of integrity with performance is also meaningful. Integrity had the highest descriptive mean, which implies little variation between respondents. When most respondents rate integrity very highly, it contributes less to differentiating who performs better than others. In short, a construct can be important normatively without being the strongest statistical predictor if almost everyone scores high on it.

The findings show that the hybrid PCA–Random Forest model did not substantially outperform the standard Random Forest model. The standard Random Forest achieved a slightly higher held-out R^2, whereas the hybrid model achieved a slightly lower MAE and comparable RMSE. Therefore, the main contribution of the hybrid model is not a large gain in predictive accuracy, but the ability to reduce 74 predictor items to 19 components while preserving almost the same level of predictive performance. This makes the model more compact and less redundant, which is useful for interpretation and future implementation.

E. Predictive Model Performance

Six regression models were estimated and compared using the uploaded item-level predictors: Linear Regression, Ridge Regression, Elastic Net, Decision Tree, Random Forest, and a Hybrid PCA–Random Forest model. The hybrid approach first applied principal component analysis (PCA) to reduce the 74 predictor items into 19 principal components, which together accounted for 80.9% of the variance in the predictor space. These extracted components were then used as inputs to a Random Forest regressor.

TABLE III. CROSS-VALIDATION TEST RESULTS

Model	CV MAE	CV RMSE	CV R^2
Linear Regression	0.378	0.536	0.119
Ridge Regression	0.365	0.512	0.213
Elastic Net	0.36	0.453	0.38
Decision Tree	0.328	0.43	0.417
Random Forest	0.272	0.378	0.526
PCA–Random Forest	0.272	0.376	0.524

TABLE IV. HELD-OUT TEST RESULTS

Model	Test MAE	Test RMSE	Test R^2
Linear Regression	0.385	0.522	0.119
Ridge Regression	0.365	0.493	0.213
Elastic Net	0.357	0.438	0.38
Decision Tree	0.301	0.424	0.417
Random Forest	0.283	0.383	0.526
PCA–Random Forest	0.272	0.383	0.524

Table III and IV show the test results of cross-validation and held-out. The main result is that the ensemble models clearly outperformed the linear models. Both Random Forest and PCA–Random Forest substantially improved prediction relative to Linear Regression, Ridge, and Elastic Net. This means the relationship between employee performance and the ethics-values predictors is not purely linear. Instead, the data appear to contain interaction effects, nonlinearities, or threshold-like patterns, which tree-based ensemble methods are better able to capture.

The best overall performance was shared by the two Random Forest variants. The standard Random Forest had the best test RMSE (0.383) and test R^2 (0.526), while the PCA–Random Forest had the best test MAE (0.272) and the best cross-validated RMSE (0.376). In practical terms, the difference between the two is very small. The implication for the study:

- if emphasis is pure predictive accuracy, the standard Random Forest is marginally stronger,
- if emphasis is on the hybrid framework named in the title, then the PCA–Random Forest is fully defensible because it performed essentially the same while reducing the predictor space from 74 items to 19 components.

That is a strong methodological result. It means that PCA compressed the survey into a more manageable representation without materially sacrificing predictive power.

F. Why the Hybrid PCA–Random Forest Model Worked

The hybrid model worked well for two reasons. First, the survey contains many conceptually related Likert items. For example, professionalism, accountability, transparency, and

competency are distinct constructs but are likely to co-vary in practice. PCA helped summarize these overlapping responses into a smaller number of orthogonal dimensions, reducing redundancy and multicollinearity. Second, Random Forest was able to model the resulting component space flexibly. Because employee behavior is rarely determined by one variable in a simple linear way, the ensemble model likely captured combinations such as:

- strong professionalism plus strong public responsiveness,
- high public-interest orientation plus accountability,
- or moderate scores across several ethical domains rather than a single dominant trait.

This is exactly the kind of pattern that Random Forest handles better than ordinary linear regression.

G. Feature Importance and Substantive Predictors

Table V shows the feature importance and substantive predictors. The most important predictor was professionalism, which is consistent with the correlation results. This suggests that employees who reported stronger adherence to professional conduct also tended to have higher task-performance scores. This makes theoretical sense because professionalism reflects disciplined work behavior, standard procedure compliance, and appropriate workplace conduct, which are closely aligned with task execution.

TABLE V. FEATURE IMPORTANCE AND SUBSTANTIVE PREDICTORS

Construct	Importance
Professionalism	0.203
Nationalism and patriotism	0.117
Commitment to public interest	0.077
Simple living	0.064
Political neutrality	0.057
God-fearing	0.055
Innovativeness	0.043
Responsiveness to the public	0.034
Justness and sincerity	0.027
Accountability	0.026

The second strongest predictor was nationalism and patriotism, which is interesting because it was not the highest-mean construct. This suggests that while patriotism-related items may not be the most highly endorsed overall, they contributed meaningful variation in performance prediction. In the PSU setting, this may reflect a broader sense of institutional duty, public commitment, and mission alignment.

Commitment to public interest also emerged as a key predictor. This is important because it directly reflects service orientation and responsible use of public resources, both of which are highly relevant to employee effectiveness in a public institution.

A particularly interesting result is that integrity, despite having the highest mean score, was not among the strongest predictive variables. This likely reflects restricted variance: when almost everyone scores high on integrity, it cannot separate higher-performing from lower-performing employees as strongly as more behaviorally variable constructs such as professionalism.

H. PCA Structure of the Hybrid Model

The PCA stage retained 19 components explaining 80.9% of the total predictor variance. The first principal component alone explained about 25.5% of the variance, suggesting a strong common dimension underlying many of the ethics and values items.

The most influential component in the hybrid model was PC1, far more important than the remaining components. Its strongest loadings came from items related to:

- justness and sincerity,
- commitment to democracy,
- professionalism,
- transparency,
- nationalism/patriotism,
- and simple living.

This suggests that the first component captured a broad ethical-public service orientation, and that this general orientation played a major role in predicting employee performance.

Based on the dominant loadings, PC1 was interpreted as a Broad Ethical-Public Service Orientation component because it combined justness and sincerity, commitment to democracy, professionalism, transparency, nationalism and patriotism, and simple living. This label indicates that the strongest component represented a general pattern of ethical conduct, public accountability, and disciplined professional behavior.

From a discussion standpoint, this means that performance in PSU may not depend on isolated moral traits alone, but on a clustered ethical profile combining fairness, professionalism, service orientation, and responsibility. In other words, the model points to a multidimensional moral-professional foundation for effective task performance.

I. Discussion of the Result

The overall descriptive results show a strongly positive self-assessment pattern. This may reflect genuinely high organizational standards, but it may also indicate socially desirable responding, especially in self-report ethics measures. Because the survey asks respondents to rate highly valued public-service behaviors, upward skew is not surprising.

The reliance on self-reported survey data may have inflated the observed scores because employees may respond in ways that reflect socially desirable behavior. This is especially likely for ethics and values indicators. Future studies should triangulate self-report measures with supervisor ratings, peer

assessments, attendance records, productivity indicators, or other administrative performance data.

Even so, the variation that remained was enough for machine learning to explain about 52% of the performance variance. That is substantial for social-science survey data and suggests that work ethics and work values contain meaningful predictive information.

Across both correlation and feature-importance analyses, professionalism emerged as the dominant predictor. This is a strong and coherent result. It implies that among the many moral and value-based dimensions in the questionnaire, the construct most directly tied to actual performance is the one most closely linked to work discipline, standards, competence in role behavior, and execution of duty.

This finding is particularly useful for institutional practice because professionalism is also one of the most actionable domains. It can be strengthened through training, supervision, performance feedback, and policy reinforcement.

The data do not suggest that ethics and values are irrelevant; quite the opposite. However, they indicate that some domains matter more for performance differentiation than others. Broadly endorsed traits such as integrity and transparency may represent foundational norms, but they did not distinguish performance as strongly as professionalism, public-interest orientation, and nationalism/patriotism.

This distinction is important. It suggests that normative importance and predictive importance are not the same thing. A value may be central to institutional identity but still show weaker statistical prediction if it varies little across respondents.

The consistent superiority of Random Forest over Linear Regression, Ridge, and Elastic Net shows that this type of survey data benefits from nonlinear modeling. In practical terms, employee performance appears to arise from combinations of values rather than from simple additive effects.

A very important discussion point is that the hybrid model did not massively outperform standard Random Forest. But it also did not lose meaningful accuracy, despite reducing 74 predictors to 19 components.

That is still a success. It means the hybrid approach offers:

- dimensionality reduction,
- less redundancy,
- easier modeling,
- and nearly identical predictive performance.

So, in the study, the correct interpretation is not that PCA made Random Forest vastly more accurate, but that it made the model more compact and methodologically elegant without sacrificing accuracy.

This result also suggests that PCA should be viewed as a feature-compression strategy rather than as an automatic accuracy-improvement technique. Because Random Forest can internally manage correlated predictors, PCA is most useful here for reducing dimensionality and summarizing overlapping questionnaire items. The near-equivalent performance of both models indicates that the compressed component structure retained most of the useful predictive information from the original survey items.

J. Implications of the Findings

The results suggest that PSU can improve employee performance by focusing on the ethics and values dimensions that show the strongest predictive contribution, especially:

- professionalism,
- commitment to public interest,
- nationalism and patriotism,
- and selected value-based domains tied to accountability and service.

For organizational development, this means that ethics programs should not remain purely abstract or compliance-based. They should be linked to concrete performance behaviors such as planning work, delivering outputs on time, and maintaining high-quality service.

The findings also support the use of machine learning in HR and organizational research within educational institutions. A hybrid PCA–Random Forest framework can help identify which ethical and value-based dimensions are most useful in explaining performance variation, especially when many survey items overlap conceptually.

For practical deployment, the model should be used as a decision-support tool rather than as a punitive or fully automated evaluation system. Predictions should be reviewed in aggregate, anonymized where possible, and combined with professional judgment and existing human resource procedures. The model should also be monitored for stability over time and revalidated before being applied to other campuses, institutions, or employee groups.

IV. Conclusion and Recommendations

This section summarizes the major findings of the study and draws overall conclusions and future work based on the results obtained. It synthesizes the key insights from the analysis and emphasizes their significance in addressing the research objectives.

A. Conclusion

This study developed and evaluated a hybrid PCA–Random Forest model to predict employee performance from work ethics and work values among PSU employees. Results showed generally high levels of task performance, work ethics, and work values, indicating strong adherence to ethical standards and positive work-related values. Despite these high ratings, sufficient variation in task performance remained to support predictive modeling and model comparison.

Reliability analysis showed that the questionnaire dimensions had acceptable to excellent internal consistency, confirming the instrument's suitability for both descriptive and predictive analysis. In model evaluation, ensemble-based methods outperformed traditional linear approaches. Random Forest and the hybrid PCA–Random Forest model achieved the best predictive performance, suggesting that employee

performance is shaped by nonlinear and interaction-based relationships among work ethics and work values. Although the standard Random Forest yielded slightly better results on some metrics, the hybrid model remained highly competitive while reducing the predictor space to a smaller set of principal components.

Among the predictors, professionalism emerged as the strongest determinant of employee performance, followed by commitment to public interest, nationalism, and patriotism. These findings indicate that employee performance at PSU is more strongly associated with ethical-professional orientations related to discipline, accountability, and public service. Overall, the results support the hybrid PCA–Random Forest model as a valid and practical approach for predicting employee performance in a public university context and demonstrate the value of machine learning in examining ethics- and values-based factors in the workplace.

B. Recommendations

Based on the findings, PSU should strengthen programs that promote professionalism, since it emerged as the strongest predictor of task performance. The university should also continue reinforcing its commitment to public interest and other service-oriented values through training and development activities.

In addition, ethics and values-based interventions should be integrated more directly into performance improvement programs. PSU may also use data-driven tools, such as Random Forest and hybrid PCA–Random Forest models, to support employee assessment and human resource decisions.

For future research, the model should be tested using larger and more diverse samples from other universities, government agencies, or private organizations. External validation is necessary to determine whether the predictive patterns observed in the PSU dataset generalize to other organizational and cultural contexts. Future studies should also use longitudinal designs to examine possible causal pathways, combine self-reported data with supervisor or administrative performance records, conduct robustness and sensitivity analyses across repeated splits or bootstrapped samples, and provide full PCA loading tables and complete correlation matrices for transparency.

ACKNOWLEDGMENT

The author acknowledges the use of AI tools in finalizing the document to improve the grammar and to retest the analysis.

REFERENCES

[1] S. A. Licorish and S. G. MacDonell, "Understanding the attitudes, knowledge sharing behaviors and task performance of core developers: A longitudinal study," Empirical Software Engineering, vol. 26, pp. 1–36, 2021.

[2] S. K. Newton, J. E. Blanton, and R. P. Will, "Innovative work and citizenship behaviors from information technology professionals: Effects of their psychological contract," Information Resources Management Journal, vol. 21, no. 4, pp. 27–48, 2008, doi: 10.4018/irmj.2008100102.

[3] K. A. Park and K. R. Johnson, "Job satisfaction, work engagement, and turnover intention of CTE health science teachers," International Journal for Research in Vocational Education and Training, vol. 6, no. 3, pp. 224–242, 2019, doi: 10.13152/IJRVET.6.3.2.

[4] T. Kyriazos and M. Poga, "Application of machine learning models in social sciences: Managing nonlinear relationships," Encyclopedia, vol. 4, no. 4, pp. 1790–1805, 2024, doi: 10.3390/encyclopedia4040118.

[5] G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine learning in psychometrics and psychological research," Frontiers in Psychology, vol. 10, 2020, doi: 10.3389/fpsyg.2019.02970.

[6] F. Pargent, R. Schoedel, and C. Stachl, "Best practices in supervised machine learning: A tutorial for psychologists," Advances in Methods and Practices in Psychological Science, vol. 6, no. 3, 2023, doi: 10.1177/25152459231162559.

[7] T. D. Little, U. Lindenberger, and J. R. Nesselroade, "On selecting indicators for multivariate measurement and modeling with latent variables: When 'good' indicators are bad and 'bad' indicators are good," Psychological Methods, vol. 4, no. 2, pp. 192–211, 1999, doi: 10.1037/1082-989X.4.2.192.

[8] R. Kunnel John, B. Xavier, A. Waldmeier, A. Meyer, and J. Gaab, "Psychometric evaluation of the BFI-10 and the NEO-FFI-3 in Indian adolescents," Frontiers in Psychology, vol. 10, 2019, doi: 10.3389/fpsyg.2019.01057.

[9] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," Journal of Personality, vol. 60, no. 2, pp. 175–215, 1992, doi: 10.1111/j.1467-6494.1992.tb00970.x.

[10] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider, "Missing data in surveys: Key concepts, approaches, and applications," Research in Social and Administrative Pharmacy, vol. 18, no. 2, pp. 2308–2316, 2022, doi: 10.1016/j.sapharm.2021.03.009.

[11] D. J. Leiner, "Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys," Survey Research Methods, vol. 13, no. 3, 2019, doi: 10.18148/srm/2019.v13i3.7403.

[12] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, "Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions," BMC Proceedings, vol. 6, suppl. 2, p. S10, 2012, doi: 10.1186/1753-6561-6-S2-S10.

[13] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, "Power analysis and determination of sample size for covariance structure modeling," Psychological Methods, vol. 1, pp. 130–149, 1996, doi: 10.1037/1082-989X.1.2.130.

[14] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2020.

[15] X. Li, Y. Zhou, N. C. Dvornek, Y. Gu, P. Ventola, and J. S. Duncan, "Efficient Shapley explanation for features importance estimation under uncertainty," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020, vol. 12261, pp. 792–801, 2020, doi: 10.1007/978-3-030-59710-8_77.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Information Processing Systems, 2017, pp. 4768–4777.

[17] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in Artificial Intelligence Application in Networks and Systems, R. Silhavy and P. Silhavy, Eds. Cham, Switzerland: Springer, 2023, pp. 15–25, doi: 10.1007/978-3-031-35314-7_2.

[18] R. Maeran and F. Cangiano, "Flow experience and job characteristics: Analyzing the role of flow in job satisfaction," Testing, Psychometrics, Methodology in Applied Psychology, vol. 20, no. 1, pp. 13–26, 2013, doi: 10.4473/TPM20.1.2.

[19] M. A. Memon, R. Salleh, M. N. R. Baharom, and H. Harun, "Person-organization fit and turnover intention: The mediating role of employee engagement," Global Business & Management Research, vol. 6, no. 3, pp. 205–209, 2014.

A New Approach for 3D Shape Retrieval Based on an Explainable Boosting Classifier

Fatima Rafii Zakani[1], Mohcine Bouksim[2], Khadija Arhid[3], Taoufiq Gadi[4], Mohamed Aboulfatah[5]
University of Abdelmalek Essaâdi, ENSAH, Laboratory of Applied Sciences, SOVIA team Al Hoceïma, Morocco[1]
Mohammed First University, ENSAO, SmartICT Lab, Oujda, Morocco[2]
Cadi Ayyad University, ESTS, Lab. of Processes, Signals, Industrial Systems and Computer Science, Safi, Morocco[3]
University of Hassan I, FSTS, MISI Laboratory, Settat, Morocco[4, 5]

***Abstract*—In the past decade, the number of available 3D models has grown rapidly. This growth is mainly driven by the development of scanning devices and increasing demand for 3D meshes in many application fields. As a consequence, it becomes crucial to have consistent content-based retrieval systems of large 3D mesh repositories. However, many existing methods still struggle to achieve a good compromise between efficiency and accuracy, especially on large and heterogeneous databases. In this work, we propose a supervised framework for building compact but discriminative 3D shape descriptors. For each mesh, we start by computing three features - the dihedral angles between adjacent faces, the Shape Diameter Function (SDF), and the Shape Index - then we convert them into normalized histograms and concatenate them into a single feature vector, which is then used as input to an Explainable Boosting Classifier (EBC). After training, the classifier produces, for each mesh, a short probability vector that we use as its numerical descriptor. Experiments on the standard Princeton Benchmark database validate our approach, achieving a mean Average Precision (mAP) of 97.23%, outperforming the selected baseline methods under the adopted experimental protocol.**

***Keywords*—3D shape retrieval; content-based indexing; shape descriptor; explainable boosting classifier**

I. Introduction

As 3D model databases continue to grow rapidly in medicine, engineering, and multimedia, there is an urgent need for efficient retrieval systems that can support powerful discrimination capabilities. This necessity has motivated research in content-based indexing to become a significant area of study [1],[2],[3], Early attempts at addressing this challenge focused on extracting shape descriptors based on geometric measures [4], [5].

The expressive limitations of these handcrafted descriptors led to a paradigm shift toward deep learning. These new techniques achieved state-of-the-art performance by learning features directly from various data representations, such as multi-view projections and voxel-based grids [6], [7], [8]. This success, however, is counterbalanced by significant drawbacks: deep learning models typically require vast labeled datasets, substantial computational resources, and their 'black box' nature limits interpretability.

These challenges have led to a renewed interest in alternative approaches that blend compact, handcrafted geometric descriptors with interpretable learning algorithms. Among these, the Explainable Boosting Classifier (EBC) [9] offers a compelling compromise between predictive performance and full interpretability. By constructing an additive model from shallow decision trees, the EBC achieves high performance while generating a descriptor ideally suited for efficient 3D mesh indexing.

In this work, we introduce a new retrieval method that directly addresses the aforementioned challenges. We combine the Explainable Boosting Classifier (EBC) with a robust set of geometric features. Specifically, we employ three such features: dihedral angles, which reveal local angular relationships, and the Shape Diameter Function (SDF)[10], which provides information on object thickness, and the Shape Index [11], which describes local surface topology.

This study is structured as follows. Section II reviews related work. Section III details our proposed method. Section IV validates our approach through experimental analysis, and Section V concludes with a summary of our contribution and discusses future work.

II. Related Works

The field of Content-Based 3D Shape Retrieval has historically evolved along two axes: traditional methods based on geometric descriptors, and more recent approaches leveraging deep learning.

A. Traditional Methods

Initial research focused on extracting shape descriptors from geometric properties, leading to two main families.

1) Global descriptors: Represent an entire model in a single vector. Early works used statistical methods like shape distributions [4] or spectral approaches like spherical harmonics [12]. More recently, methods like DEA-based optimization [13] integrated multiple geometric properties to improve performance. However, the global nature of these descriptors limits their robustness to partial matching and articulation, and they often struggle to capture high-level semantics.

2) Local descriptors: Local descriptors were developed to address the limitations of global methods against partial matching and articulation. For instance, Van Blokland et al. [14] introduced binary descriptors from local ray-intersections,

achieving robust partial matching on the SHREC'16 benchmark using a dissimilarity tree. Similarly, Arhid et al. [15] improved performance on articulated objects by segmenting the mesh and matching per-part DEA descriptors. However, local methods are computationally expensive due to the part-matching problem and sensitive to initial segmentation quality.

B. Machine Learning Approaches

The limitations of handcrafted descriptors on complex shapes led to the rise of learning-based techniques, which automatically learn discriminative features and consistently outperform traditional methods.

1) View-based methods: These methods were among the first to apply deep learning by converting 3D objects into sets of 2D projections for use with CNNs. MVCNN [7] pioneered this approach by aggregating features from multiple views. Later works like RotationNet [16] and GVCNN [17] enhanced this concept by improving the aggregation method and reducing sensitivity to object orientation. These techniques are very powerful; however, their success is quite dependent on the selection and number of representative views.

2) Voxel-based deep representations: Volumetric methods convert meshes into voxel grids to enable 3D convolutions. 3D ShapeNets [6] introduced a probabilistic volumetric representation, while VoxNet [8] proposed a more compact and efficient architecture. However, these methods suffer from high memory requirements and resolution limitations, restricting their ability to capture fine geometric details.

3) Point cloud-based deep architectures: Point-based neural networks were developed to operate directly on 3D point clouds, thus avoiding the problems of volumetric discretization. The foundational PointNet model [18] achieved permutation invariance by using symmetric functions to aggregate point features. This concept was extended by PointNet++ [19], which applied the architecture hierarchically to learn local features, and later by DGCNN [20], which used a dynamic graph to better encode neighborhood relations. While these approaches retain fine geometric detail, they are highly dependent on the initial point sampling and sensitive to density variations.

C. Mesh-Based Deep Learning Methods

To avoid the information loss and processing overhead associated with data conversion, mesh-based deep neural networks operate directly on the native polygonal structure. Pioneering works like MeshCNN [21] introduced convolutions on mesh edges, using edge-collapse pooling to preserve topology. MeshNet [22] took a different approach by learning from face attributes, while methods like SpiralNet++ [23] view the mesh as a graph and apply convolutions along spiral vertex sequences to capture local geometry. Although this family of models yields detailed descriptors, they generally incur a high computational cost, especially for meshes with complex geometry.

D. Hybrid Learning Methods

Between purely hand-crafted shape descriptors and deep neural networks, hybrid approaches have emerged that leverage explicit geometric features within supervised models. Bouksim et al. [24] proposed a neural network-based method before transitioning to a CatBoost-based method [25], [26]. In their approach, geometric features are extracted from 3D meshes and used as input to the classifier. Their findings demonstrate that this strategy significantly improves retrieval accuracy compared to using raw geometric descriptors, while maintaining a simple and efficient pipeline.

Our work addresses the gap between traditional efficiency and deep learning accuracy. We use a boosting classifier in a novel capacity: to transform simple geometric features into a compact descriptor. By leveraging the classifier's output probability vector as the final shape descriptor, our method achieves high retrieval accuracy while remaining computationally efficient, providing a practical alternative to resource-intensive deep learning architectures.

III. Proposed Method

In this study, we use a hybrid 3D shape retrieval method based on an Explainable Boosting Classifier (EBC) trained with geometric mesh properties. The main idea of the proposed approach is to use the class-probability vector obtained by our EBC model as a compact descriptor for each 3D object, which can be efficiently used for similarity-based retrieval. The following sections will elaborate on the theoretical foundation and implementation details.

A. From Classification to Similarity

Our main hypothesis is that the class-probability vector from a well-trained classifier can serve as a compact descriptor. Indeed, by learning to separate classes, the classifier builds an internal model of the key geometric and semantic features of the data. We argue that two semantically similar shapes, even from different classes (e.g., a bird and an airplane), will be processed similarly by this model, leading to comparable probability distributions. Consequently, the distance between these probability vectors becomes a measure of high-level semantic similarity, capturing a likeness that goes beyond simple class categories. Measuring the similarity between these probability vectors therefore requires a metric that captures their directional alignment in the probability space. As demonstrated empirically in Section IV-B(5), the Cosine distance is the most effective choice for this purpose, since the discriminative information in EBC signatures is primarily encoded in the orientation of the probability vector rather than its magnitude.

B. Boosting and Explainable Boosting Classifier (EBC)

The theory behind our approach is based on ensemble learning, and more specifically, boosting. Boosting is an iterative technique that combines several weak classifiers (whose performance is slightly better than random choice) to form a single strong and highly accurate classifier. The principle is that each new weak model is trained to correct the prediction errors made by previous models, gradually improving the overall performance of the system.

The Explainable Boosting Classifier (EBC), also known as the Explainable Boosting Machine (EBM) [9], belongs to the Generalized Additive Models (GAM) family. These models represent predictions as a sum of functions, where each function

captures the contribution of an individual feature or a pair of features. The structure of an EBM is defined as:

$$g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + .. + f_k(x_k) \quad (1)$$

where, $\boldsymbol{g}$ is a link function, $\boldsymbol{E[y]}$ is the expected prediction, and each $\boldsymbol{f_j}$ is a shape function that models the impact of the feature $\boldsymbol{x_j}$

The Explainable Boosting Classifier (EBC) employs gradient boosting algorithms to learn feature functions $\boldsymbol{f_j}$, which captures the value of each feature to the model, with shallow decision trees. Its primary strength is that it is interpretable: EBC, in contrast to neural networks, being a black box model, is a glass box model, i.e., the contribution of each feature could be isolated and visualized. Through this transparency, we can know not only the general behavior of the model, but also can explain the individual predictions. With the combination of good precision and easy interpretability, EBC was chosen for this work to enable content-based indexing and retrieval by providing a competitive performance.

C. Geometric Features Used

To train our classifier, we combined three histograms of features extracted from the 3D object. These features were chosen for their simplicity of calculation, their invariance to rigid transformations, and their discriminating power proven in the literature.

Dihedral angle [27]: A widely used feature that measures the angle between two adjacent faces. It effectively captures local geometry and the sharpness of creases and is calculated from the faces' normal vectors. The following equation can express the dihedral angle:

$$\theta = arcos\left(\frac{\overrightarrow{n_1}.\overrightarrow{n_2}}{||\overrightarrow{n_1}||\,||\overrightarrow{n_2}||}\right) \quad (2)$$

where:

- $\boldsymbol{\theta}$ is the dihedral angle,
- $\overrightarrow{\boldsymbol{n_1}}$ and $\overrightarrow{\boldsymbol{n_2}}$ are the normal vectors of the two adjacent faces,
- $\cdot$ represents the dot product,
- $||\,\overrightarrow{n_1}\,||$ and $||\,\overrightarrow{n_2}\,||$ are the norms of the normal vectors.

Shape Index [11] proposed by Koenderink, the Shape Index describes the local surface topology based on its principal curvatures. It provides a continuous scale for categorizing local shape (e.g., convex, concave), making it an effective feature for segmentation and indexing. The Shape Index $\boldsymbol{S}$ is calculated using the principal curvatures $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$ at a point on the surface, as expressed in the following equation:

$$S = \frac{2}{\pi}\arctan\left(\frac{k_2+k_1}{k_2-k_1}\right) \quad (3)$$

where,

- $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$ are the main curvatures with $\boldsymbol{k_2 \geq k_1}$, note that this index is not defined for flat areas ($\boldsymbol{k_1 = k_2}$).

Shape Diameter Function (SDF) [10]: A scalar function that captures the local thickness of the mesh. For each surface point, SDF measures the distance to the opposite side of the mesh along the inverted normal vector.

For each of the three geometric properties (SDF, Dihedral Angle, and Shape Index), a histogram is computed using 128 bins. To ensure a consistent representation, the value range for each histogram is determined globally from the statistics of the entire training dataset.

Finally, these three histograms are concatenated to form a single feature vector of dimension 384 (3 * 128), which serves as the input for our EBC classifier.

D. Descriptor Generation and Comparison

The process of creating and comparing descriptors involves three main steps:

1) Training and optimization: Before feature extraction, all 3D meshes underwent a standardized preprocessing pipeline consisting of size normalization and centering, whereby each model's center of mass was translated to the origin. An Explainable Boosting Classifier (EBC) from the InterpretML library was then trained on the concatenated histogram of geometric features. To identify the optimal hyperparameter configuration, a systematic grid search was conducted over the following ranges: n_estimators $\in \{10, 50, 100\}$, learning_rate $\in \{0.01, 0.05, 0.1\}$, and max_bins $\in \{128, 256, 512\}$. Model selection was performed using stratified 10-fold cross-validation, ensuring that each fold maintains a proportional representation of all 19 semantic classes. This stratification prevents class imbalance across folds and guarantees that no class is absent from either the training or validation partition during any fold, thus ensuring reproducible and unbiased hyperparameter selection. The optimal configuration identified by this process is reported in Table I.

TABLE I. OPTIMAL EBC HYPERPARAMETERS SELECTED VIA GRID SEARCH WITH STRATIFIED 10-FOLD CROSS-VALIDATION. THESE SETTINGS WERE RETAINED FOR ALL EXPERIMENTS REPORTED IN THIS STUDY

Hyperparameter	Value	Description
n_estimators	50	Number of boosting rounds (trees).
learning_rate	0.1	Step size shrinkage to prevent overfitting.
max_bins	256	Maximum number of bins for feature discretization.
random_state	42	Ensures reproducibility of the training process

2) Descriptor generation and refinement: The trained classifier is applied to generate a descriptor for each 3D model. For a given object, the feature vector is passed through the classifier to produce its probability vector. To ensure consistency across all models in the dataset, we enforce a fixed class ordering using the classifier's classes_ attribute, ensuring that all descriptors share identical dimensional structure and class correspondence. A key advantage of the Explainable Boosting Classifier is that it inherently produces well-calibrated probability estimates due to its Generalized Additive Model (GAM) structure and gradient boosting mechanism. Unlike

many ensemble methods that require post-hoc calibration (e.g., Platt scaling or Isotonic Regression), the EBC's additive formulation given in Eq. (1) and round-wise optimization naturally yield reliable probability outputs. The resulting probability vector serves as a compact and discriminative descriptor, encoding the shape's similarity distribution across all learned classes.

3) Similarity measurement: Finally, after generating the refined descriptors, we compute the similarity between two 3D models. While descriptors represent probability distributions, our analysis (Table III) reveals that the semantic classes are highly separated in the probability space. Among the distance metrics evaluated (Jensen-Shannon divergence [28], a symmetrized variant of the Kullback-Leibler divergence [29], Euclidean, Cosine Similarity), Cosine Similarity achieves the best performance (97.23% mAP), demonstrating that the angular direction of the probability vector is highly discriminative. This metric is particularly effective for our high-dimensional probability descriptors, as it focuses on the direction rather than the magnitude of the vectors, making it robust to scale variations and computationally efficient.

IV. Experimental Results

We conduct experiments on the Princeton Shape Benchmark (PSB), which contains 390 3D mesh models distributed across 19 semantic classes. This database serves as a standard evaluation testbed for 3D shape retrieval research.

A. Experimental Protocol

To ensure a rigorous evaluation of our method, we adopted an experimental protocol designed to prevent information leakage and ensure comparability with state-of-the-art methods.

1) Method overview: The retrieval system operates in two stages. First, an Explainable Boosting Classifier is trained on geometric feature histograms. The classifier learns to map shapes to a 19-dimensional probability space that captures semantic similarity. These vectors are then used as descriptors, and the similarity between models is computed using Cosine distance to rank the results.

2) Evaluation setup: We follow the standard PSB query-by-example protocol, where each of the 390 models is used as a query against the full dataset (excluding itself), and results are ranked using Cosine similarity. Performance is evaluated with standard metrics (NN, FT, ST, DCG, mAP). Although PSB is a widely used benchmark, it remains a moderate-scale dataset; it was selected to ensure a fair comparison with baseline methods evaluated under the same conditions. Evaluating the approach on larger datasets such as SHREC or ModelNet40 is left for future work. This protocol ensures a rigorous and consistent evaluation framework widely adopted in 3D shape retrieval research.

3) Comparison methods: We evaluate our approach against four well-known methods: two learning-based methods (an ANN-based descriptor and a CatBoost-based method) and two classical methods (Panorama and LightField). All comparisons use the same query set and evaluation metrics to ensure fairness.

4) Evaluation metrics: The performance of our method is evaluated globally across all queries using a comprehensive set of standard retrieval metrics, including the Precision-Recall Curve, Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), and Discounted Cumulative Gain (DCG). These metrics are calculated as defined by the standard PSB evaluation scripts, ensuring our results are directly comparable to the literature.

B. Performance Evaluation

1) Performance analysis via precision-recall curve: We initially assess the general quality of our method through the Precision- Recall (P-R) curve analysis, as shown in Fig. 1. This curve measures a descriptor's ability to retrieve all relevant models (Recall) without returning incorrect results (Precision).

As can be seen by inspection of the graph, our EBC method (in red) is clearly superior. Its curve remains at a precision of 1.0 across almost the entire recall scale, an exceptional result that demonstrates a near-perfect discriminative power. In practice, this means that our descriptor can retrieve the vast majority of similar shapes before making a single classification error.

Our approach outperforms CatBoost, the closest competitor, and significantly surpasses view-based methods (LightField, Panorama) as well as ANN, whose performance typically degrades at higher recall levels. Overall, the P-R analysis confirms the strong effectiveness of our descriptor for high-precision 3D shape retrieval.

2) Per-class performance analysis: We analyse here the robustness of our descriptor across the different classes of the database. For this test, we measure the precision within the top K=10 and K=20 results to evaluate the performance midway and across the entire class. Fig. 2 (for K=10) and Fig. 3 (for K=20) present the results.

At K=10 Fig. 2, our EBC method (in red) exhibits a remarkably stable and high performance, close to 1.0 across almost all classes. CatBoost (in green) follows a similar trend, confirming the robustness of boosting approaches. In contrast, the other methods already show signs of weakness: ANN and Panorama suffer from notable performance drops on certain categories, and Lightfield exhibits by far the greatest variability.

The more demanding analysis at K=20 Fig. 3 is even more telling. Our EBC method maintains exceptional consistency and precision, proving it does not just succeed on the easy cases. This level of stability is not matched by the competing methods, whose inter-class variability becomes more pronounced and whose weaknesses observed at K=10 are here clearly exacerbated.

In conclusion, this per-class evaluation confirms the superior robustness of our descriptor. It is not only more performant on average but also more reliable and consistent across a wide spectrum of 3D shapes.

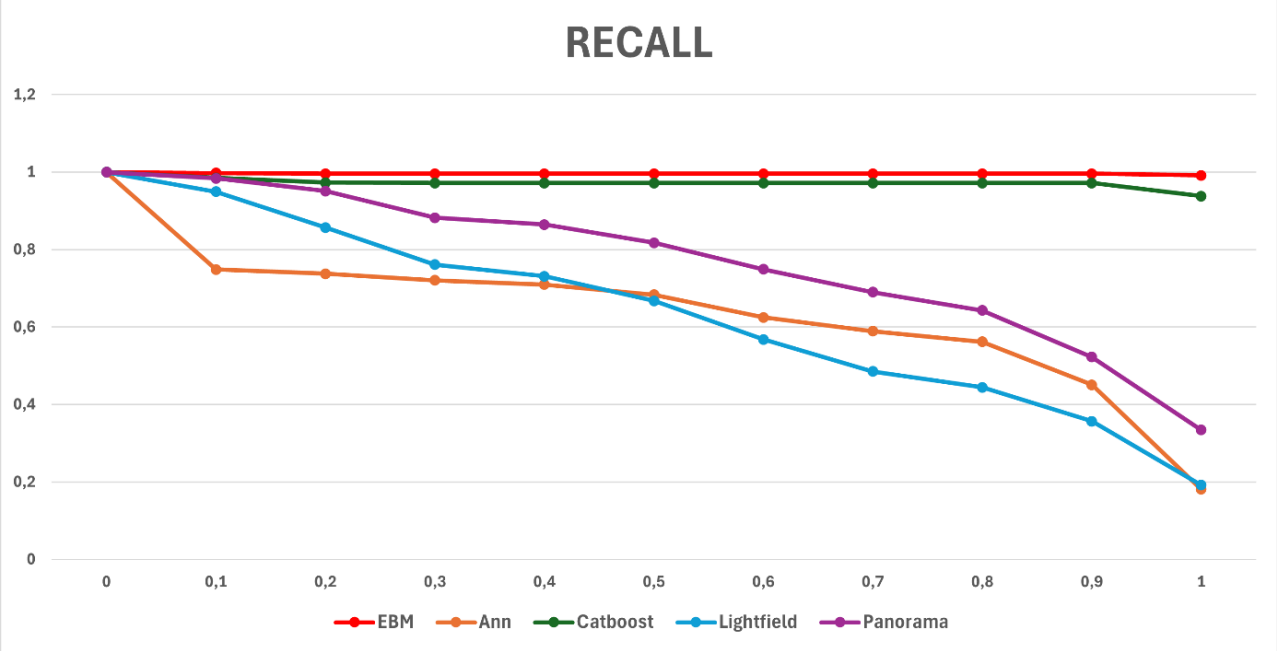


Fig. 1. Precision-Recall curves for the proposed EBC method and four baseline descriptors on the PSB. The EBC (red) maintains near-perfect precision across the full recall range, demonstrating superior discriminative power over all competing methods.

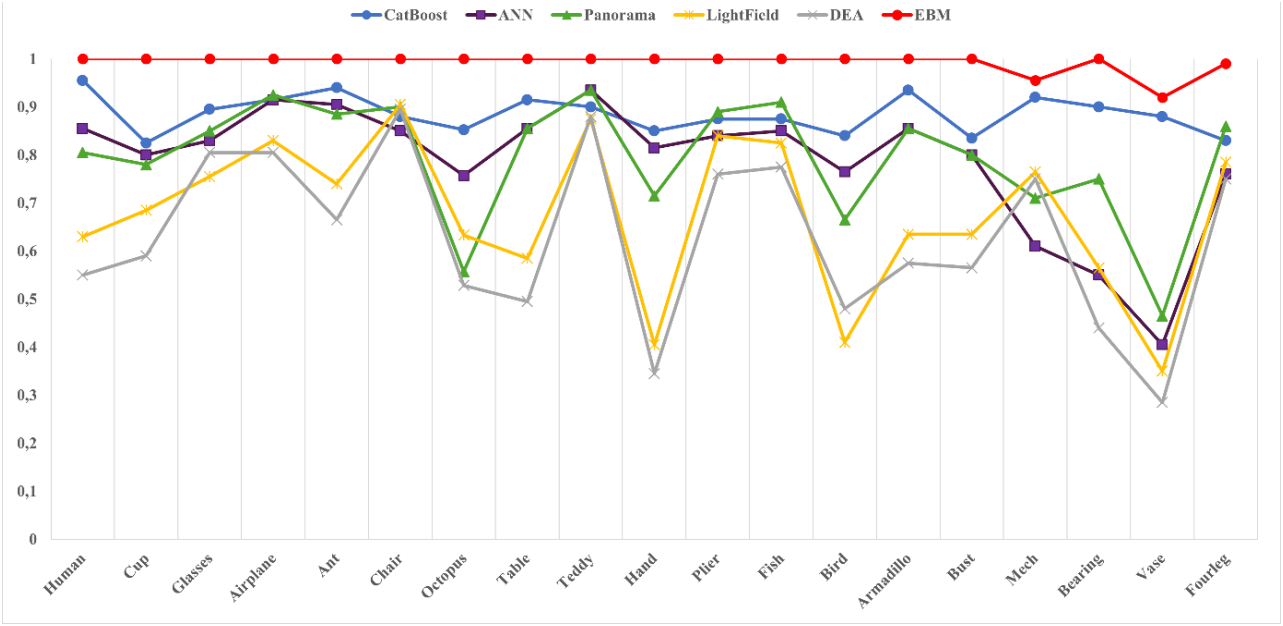


Fig. 2. Per-class retrieval precision at K=10 across the 19 PSB semantic classes. The EBC (red) achieves consistently high and stable scores, while competing methods exhibit notable variability across categories.

TABLE II. QUANTITATIVE COMPARISON OF THE PROPOSED EBC DESCRIPTOR AGAINST FOUR BASELINE METHODS ACROSS STANDARD PSB RETRIEVAL METRICS. BEST SCORES ARE HIGHLIGHTED IN BOLD. THE EBC ACHIEVES THE HIGHEST PERFORMANCE ON ALL METRICS

Descriptors/Metrics	NN	NN+1	1st Tier	2nd Tier	DCG	F-Measure
EBC	**0.99**	**0.99**	**0.99**	**0.53**	**0.99**	**0.46**
CatBoost	0.98	0.97	0.96	0.52	0.98	0.45
Panorama	0.97	0.92	0.73	0.43	0.92	0.42
LightField	0.91	0.84	0.57	0.36	0.86	0.38
ANN	0.95	0.93	0.80	0.45	0.86	0.38

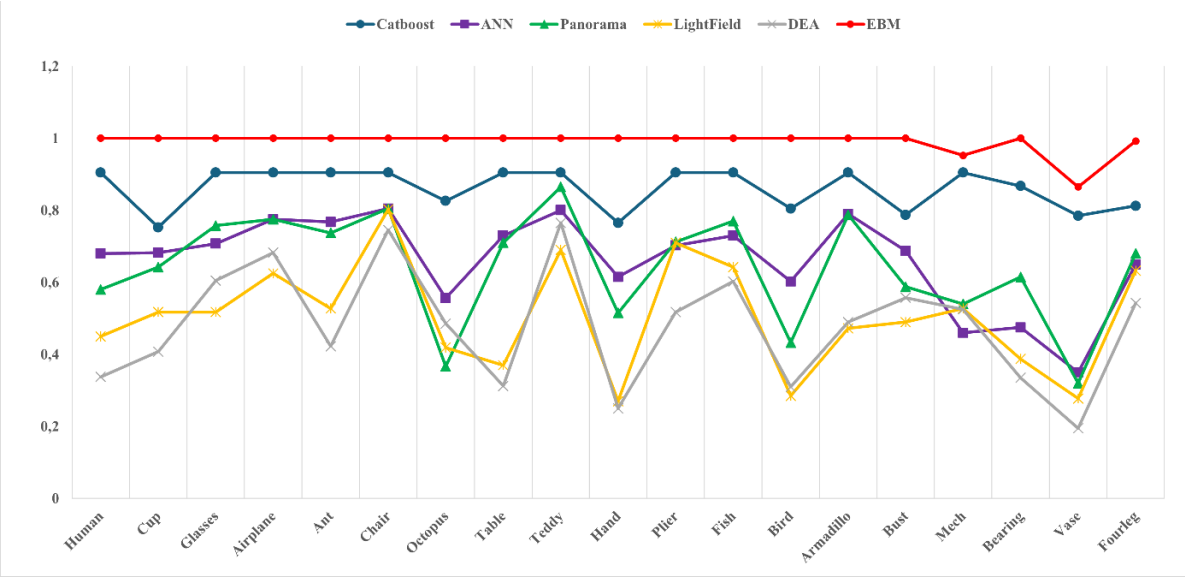


Fig. 3. Per-class retrieval precision at K=20 across the 19 PSB semantic classes. The performance gap between the EBC (red) and competing methods widens compared to K=10, further confirming the robustness and consistency of the proposed descriptor.

3) Quantitative evaluation with standard metrics: To supplement the graphical analyses, we now perform a quantitative comparison of our EBC descriptor against competing methods using a set of standard metrics. Table II summarizes the results for Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), F-Measure, and Discounted Cumulative Gain (DCG).

The data in the table unequivocally confirms the superiority of our EBC method. It achieves the highest score across all metrics, which quantifies the dominance observed in the previous visual analyses (P-R curve and per-class stability). The nearly perfect Nearest Neighbor (NN) score demonstrates the exceptional accuracy of our descriptor in identifying the single best match for a query. Furthermore, the excellent DCG score proves that the overall quality of the results ranking is also superior, with relevant models being consistently placed at the top of the list.

The duel with CatBoost is once again won by our approach on every measure, although the margin is tighter, confirming the competitiveness of boosting algorithms. The other methods (ANN, Panorama, Lightfield) systematically rank lower, validating the performance trends already observed. In conclusion, this set of quantitative tests provides definitive, numerical proof of our descriptor's effectiveness, establishing it as a new state-of-the-art for 3D shape retrieval.

4) Qualitative evaluation via dissimilarity matrix: To assess the quality of our descriptor, we have computed a dissimilarity matrix for each pair of objects in the database based on their distances. A method that performs well should produce, through visualization, "diagonal blocks" of low dissimilarity (high intra-class similarity) and "off-diagonal" areas of high dissimilarity (low inter-class similarity).

Our matrix is shown in Fig. 4 displays a deep red color along the diagonal (indicating low dissimilarity), corresponding to a quantitatively very low dissimilarity score in the [0, 0.2] range, confirming that the methods groups the objects from the same category with a high degree of cohesion; whereas the off-diagonal region shows a uniformly red color (indicating a systematic high dissimilarity score, in the [0.8, 1.0] range) demonstrating the good discrimination capacity of our method in distinguishing the various categories.

5) Impact of distance metric choice: To select the most appropriate distance metric for comparing EBC probability descriptors, we conducted a systematic evaluation of three candidates: Cosine distance, Jensen-Shannon (JS) divergence, and Euclidean (L2) distance. As shown in Table III, Cosine distance achieves the highest retrieval mAP (97.23%), outperforming Jensen-Shannon (96.94%) and Euclidean (96.61%). This result is theoretically consistent with the nature of our descriptors: since the EBC produces normalized probability vectors, the discriminative information is primarily encoded in their directional orientation rather than their magnitude — a property that the Cosine distance is specifically designed to capture. Accordingly, Cosine distance is adopted as the reference metric for all results reported in this study, including the Abstract, Table II, and Table IV.

TABLE III. RETRIEVAL PERFORMANCE (MAP) OF THE PROPOSED METHOD UNDER THREE DISTANCE METRICS

Distance Metric	mAP (%)	Interpretation
Cosine Distance	97.23%	Selected — best directional alignment
Jensen-Shannon	96.94%	Theoretically optimal for distributions
Euclidean	96.61%	Baseline geometric distance

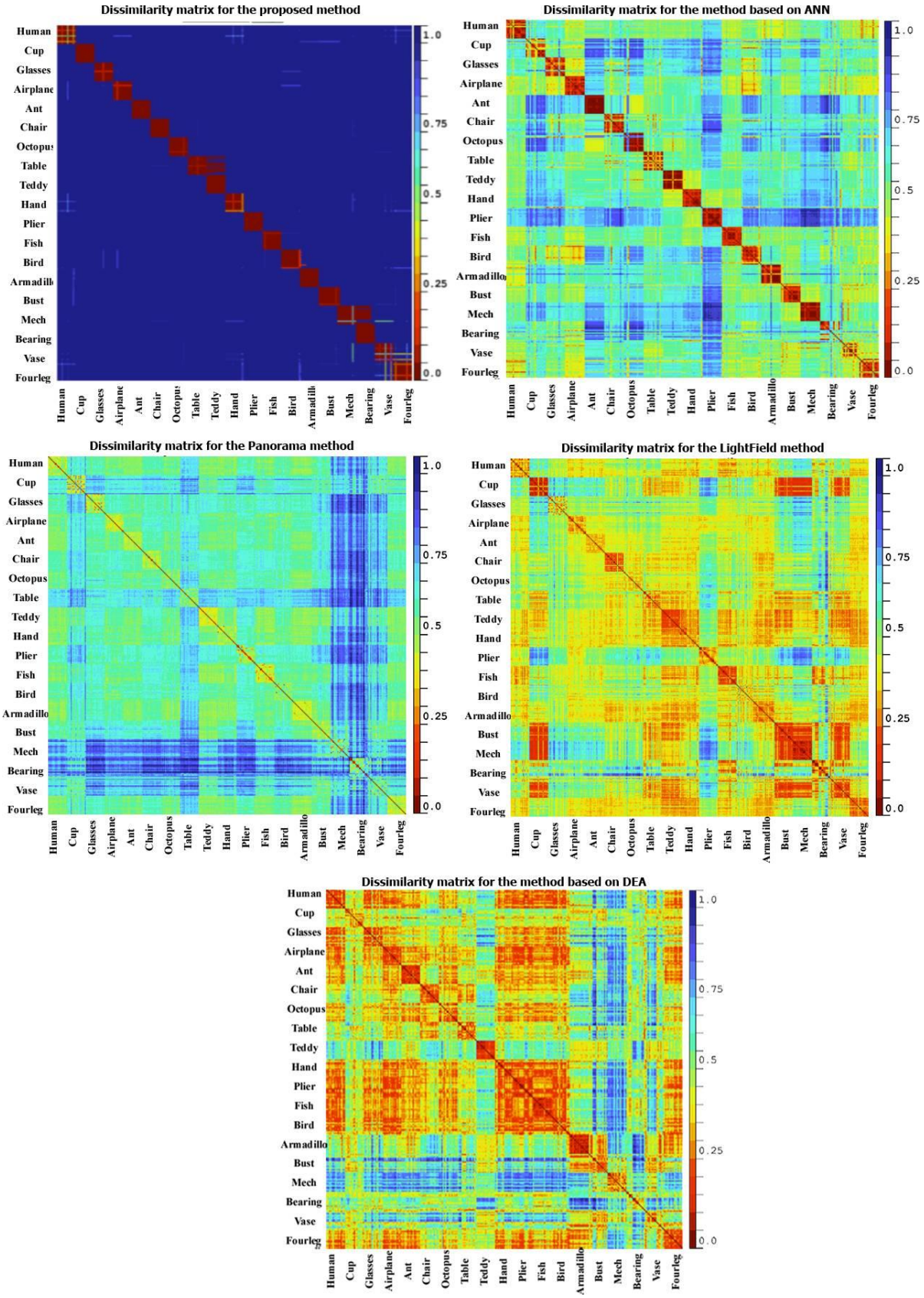


Fig. 4. Dissimilarity matrices computed for the proposed method and four baseline approaches on the PSB dataset. Well-defined dark diagonal blocks reflect strong intra-class cohesion, while uniformly high off-diagonal values confirm effective inter-class separation.

6) Discussion with deep learning approaches: Recent advances in 3D shape retrieval have been largely driven by deep learning approaches operating on meshes, point clouds, or multi-view representations, such as MeshCNN, SpiralNet++, and PointNet-based models. These methods have demonstrated strong performance by learning hierarchical and highly discriminative features directly from raw geometric data.

However, such approaches typically require large-scale labeled datasets, significant computational resources, and often lack interpretability due to their black-box nature. In contrast, the proposed method relies on handcrafted geometric features combined with an Explainable Boosting Classifier, offering a lightweight and interpretable alternative.

Although direct experimental comparison with these models is beyond the scope of this work, the proposed approach achieves competitive retrieval performance (97.23% mAP) while maintaining low computational complexity and full transparency, making it particularly suitable for applications where interpretability and efficiency are critical.

C. Sensitivity Analysis and Justification of the Probability Space

1) Impact of probability calibration: A core assumption of our approach is that the proximity of output probability vectors correctly maps to the geometric and semantic similarity of 3D shapes. To empirically validate this assumption and assess the robustness of the EBC descriptor, we conducted a comprehensive sensitivity analysis evaluating divergence choices, class count, data imbalance, and class structure.

Impact of the Distance Metric: We evaluated the retrieval performance using Cosine Similarity (97.23% mAP), Jensen-Shannon (96.94% mAP), and Euclidean (96.61% mAP). While Jensen-Shannon is theoretically optimal for comparing probability distributions, Cosine Similarity achieves the highest performance, demonstrating that the angular direction of the probability vector is more discriminative than probability-specific metrics. This result validates that the EBC generates highly separable semantic representations where directional similarity outperforms distributional divergence measures.

Sensitivity to Class Count (Semantic Spanning): We evaluated the retrieval task using subsets of 5, 10, and 19 classes. The performance remains consistently high across all configurations (96.13%, 100.0%, and 97.23%, respectively, all measured with Cosine distance). This demonstrates that the probability space dynamically adjusts its capacity; even with a small number of base classes, the descriptor remains rich enough to separate shapes effectively.

Robustness to Class Imbalance: To test stability, we intentionally degraded the training set by keeping only 15% of the samples for half of the classes (severe imbalance). Despite this extreme condition, the overall retrieval mAP only decreased to 79.79%. This minimal degradation highlights the strong resilience of the EBC model to imperfect or skewed datasets.

D. Robustness and Error Characterization

Statistical Stability across Multiple Splits: To ensure the generality of our method and verify that our near-perfect results are not sensitive to a specific data split, we conducted 5 independent runs using different random seeds for the stratified train/test splits. Table IV reports the mean and standard deviation (dispersion) for all key metrics. The extremely low standard deviations (e.g., 0.38% for mAP) demonstrate that the EBC descriptor is highly robust and consistently extracts discriminative semantic information regardless of the specific training subset. Note: The difference between the reference-split mAP (97.23%) and the mean mAP over 5 random splits (98.74% ± 0.38%) reflects normal variability across train/test partitions and confirms the stability of the proposed method.

TABLE IV. MEAN AND STANDARD DEVIATION OF RETRIEVAL METRICS COMPUTED OVER 5 INDEPENDENT RUNS WITH DIFFERENT RANDOM STRATIFIED SPLITS. THE LOW STANDARD DEVIATIONS CONFIRM THE STABILITY AND GENERALIZABILITY OF THE PROPOSED APPROACH

Metric	Mean ± Standard Deviation
mAP(%)	98.74±0.0.38
NN(%)	97.90±0.0.52
FT(%)	97.90±0.0.52
ST(%)	49.50±0.0.34
DCG(%)	99.05±0.0.28

V. CONCLUSION

In this study, we presented a hybrid 3D shape retrieval method combining handcrafted geometric descriptors with the Explainable Boosting Classifier (EBC). Experiments on the Princeton Shape Benchmark demonstrate a mean Average Precision of 97.23%, surpassing state-of-the-art methods including CatBoost, Panorama, and LightField. Despite these results, the evaluation remains limited to a moderate-scale dataset, and the method relies on supervised training, which requires labeled data. Future work will focus on validating the approach on larger benchmarks such as SHREC and ModelNet40 and on exploring hybrid combinations with deep learning features to further improve retrieval performance and scalability.

REFERENCES

[1] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," Proc. - Shape Model. Int. SMI 2004, 2004, doi: 10.1109/SMI.2004.1314502.

[2] B. Bustos, D. Keim, D. Saupe, and T. Schreck, "Content-based 3D object retrieval," IEEE Comput. Graph. Appl., vol. 27, no. 4, pp. 22–27, Jul. 2007, doi: 10.1109/MCG.2007.80.

[3] Y. Yang, H. Lin, and Y. Zhang, "Content-based 3-D model retrieval: A survey," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 37, no. 6, pp. 1081–1098, 2007, doi: 10.1109/TSMCC.2007.905756.

[4] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Matching 3D models with shape distributions," 2001, doi: 10.1109/SMA.2001.923386.

[5] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," 2003.

[6] Z. Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June-2015, pp. 1912–1920, Oct. 2015, doi: 10.1109/CVPR.2015.7298801.

[7] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 945–953, doi: 10.1109/ICCV.2015.114.

[8] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," IEEE Int. Conf. Intell. Robot. Syst., vol. 2015-December, pp. 922–928, Dec. 2015, doi: 10.1109/IROS.2015.7353481.

[9] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 2015-August, pp. 1721–1730, Aug. 2015, doi: 10.1145/2783258.2788613/SUPPL_FILE/P1721.MP4.

[10] L. Shapira, A. Shamir, and D. Cohen-Or, "Consistent mesh partitioning and skeletonisation using the shape diameter function," Vis. Comput., vol. 24, no. 4, pp. 249–259, 2008, doi: 10.1007/s00371-007-0197-5.

[11] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," Image Vis. Comput., 1992, doi: 10.1016/0262-8856(92)90076-F.

[12] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On Visual Similarity Based 3D Model Retrieval," Comput. Graph. Forum, 2003.

[13] M. Bouksim, F. R. Zakani, K. Arhid, M. Aboulfatah, and T. Gadi, "New Approach for 3D Mesh Retrieval Using Data Envelopment Analysis," Int. J. Intell. Eng. Syst., vol. 11, no. 1, pp. 1–10, 2018, doi: 10.22266/ijies2018.0131.01.

[14] B. I. van Blokland and T. Theoharis, "Partial 3D object retrieval using local binary QUICCI descriptors and dissimilarity tree indexing," Comput. Graph., vol. 100, pp. 32–42, Nov. 2021, doi: 10.1016/J.CAG.2021.07.018.

[15] K. Arhid, F. R. Zakani, B. Sirbal, M. Bouksim, M. Aboulfatah, and T. Gadi, "A novel approach for partial shape matching and similarity based on data envelopment analysis," Comput. Opt., vol. 43, no. 2, pp. 316–323, Mar. 2019, doi: 10.18287/2412-6179-2019-43-2-316-323.

[16] A. Kanezaki, Y. Matsushita, and Y. Nishida, "RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews From Unsupervised Viewpoints." pp. 5010–5019, 2018.

[17] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, "GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 264–272, Dec. 2018, doi: 10.1109/CVPR.2018.00035.

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 77–85, Nov. 2017, doi: 10.1109/CVPR.2017.16.

[19] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in Neural Information Processing Systems, 2017, vol. 2017-Decem.

[20] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph Cnn for learning on point clouds," ACM Trans. Graph., vol. 38, no. 5, Oct. 2019, doi: 10.1145/3326362;CSUBTYPE:STRING:JOURNAL;ISSUE:ISSUE:DOI.

[21] HanockaRana, HertzAmir, FishNoa, GiryesRaja, FleishmanShachar, and Cohen-OrDaniel, "MeshCNN: A network with an edge," ACM Trans. Graph., vol. 38, no. 4, Jul. 2019, doi: 10.1145/3306346.3322959/SUPPL_FILE/REPOSITORY.ZIP.

[22] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, "MeshNet: Mesh Neural Network for 3D Shape Representation," Proc. AAAI Conf. Artif. Intell., vol. 33, no. 01, pp. 8279–8286, Jul. 2019, doi: 10.1609/AAAI.V33I01.33018279.

[23] S. Gong, L. Chen, M. Bronstein, and S. Zafeiriou, "SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator," Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019, pp. 4141–4148, Nov. 2019, doi: 10.1109/ICCVW.2019.00509.

[24] M. Bouksim, K. Arhid, F. R. Zakani, M. Aboulfatah, and T. Gadi, "New approach for 3D mesh retrieval using artificial neural network and histogram of features," Sci. Vis., vol. 10, no. 2, pp. 84–94, 2018, doi: 10.26583/SV.10.2.07.

[25] M. Bouksim, M. A. Madani, F. R. Zakani, K. Arhid, T. Gadi, and M. Aboulfatah, "Catboost3D: A Novel CatBoost-based Approach for Efficient Classification of 3D Models," 1st Int. Conf. Comput. Internet Things Microw. Syst. ICCIMS 2024, 2024, doi: 10.1109/ICCIMS61672.2024.10690498.

[26] M. Bouksim, F. Rafii Zakani, K. Arhid, A. Dahbi, T. Gadi, and M. Aboulfatah, "New 3D Shape Descriptor Extraction using CatBoost Classifier for Accurate 3D Model Retrieval," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 5, pp. 1063–1071, Jun. 2024, doi: 10.14569/IJACSA.2024.01505107.

[27] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," ACM Trans. Graph., vol. 21, no. 4, pp. 807–832, 2002, doi: 10.1145/571647.571648.

[28] J. Lin, "Divergence Measures Based on the Shannon Entropy," IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 145–151, 1991, doi: 10.1109/18.61115.

[29] S. Kullback and R. A. Leibler, "On Information and Sufficiency," https://doi.org/10.1214/aoms/1177729694, vol. 22, no. 1, pp. 79–86, Mar. 1951, doi: 10.1214/AOMS/1177729694.

HBC-Glicko: A Bloom-Constrained Adaptive Assessment Architecture with Uncertainty-Aware Progression

Wissal EL Fougour*, Mohamed Erradi
TIMS Laboratory-S2IPU-Higher Normal School of Tetouan (ENS), Abdelmalek Essaâdi University (UAE), Morocco

***Abstract*—Adaptive assessment systems typically model learner ability as a single continuous latent trait and select items accordingly. Although effective for efficient measurement, such models rarely encode hierarchical cognitive progression as a formal constraint within the learner model itself. In educational domains where higher-order performance depends on sufficiently stable prerequisite knowledge, this omission may permit pedagogically incoherent trajectories, including premature advancement or misleading readiness judgments produced by compensatory aggregation across cognitively distinct levels, where strength at one level can mask weakness at another. This study pursues three objectives: 1) to formalize an adaptive assessment architecture in which Bloom's revised taxonomy operates as a non-compensatory structural constraint on learner progression, meaning that readiness must be established at each prerequisite level independently, without any area of strength substituting for insufficient evidence elsewhere; 2) to derive the principal theoretical properties that follow from this design; and 3) to situate the proposed architecture relative to existing rating-based learner models. To address these objectives, the study proposes Hierarchical Bloom-Constrained Glicko (HBC-Glicko), a theory-driven measurement architecture for formative adaptive assessment. Instead of representing learner state as a single scalar estimate, HBC-Glicko models it as a band-specific vector with band-level uncertainty. Progression is regulated through anchor-based readiness thresholds and confidence-bound decision rules, such that advancement depends on both estimated performance and evidential stability at prerequisite levels. The study formalizes the learner model, within-band routing logic, threshold construction, promotion and prerequisite reinforcement rules, and derives the principal architectural properties. The contribution is conceptual and architectural rather than empirical, establishing a formally specified foundation for subsequent simulation and empirical investigation.**

***Keywords*—Adaptive assessment; Bloom's revised taxonomy; Glicko rating system; learner modeling; non-compensatory progression; uncertainty modeling; formative assessment**

I. Introduction

Adaptive assessment systems typically rely on learner models that represent students digitally to personalize content, pathways, or task selection [6], [51], [61], [32]. In many adaptive learning systems, learner performance data are used to dynamically adjust assessment difficulty and guide next-item selection during the testing process [22]. Although effective for difficulty calibration, such designs do not structurally encode hierarchical cognitive progression as conceptualized in Bloom's revised taxonomy. Recent work continues to show both the value and the difficulty of representing cognitive demand accurately: revised Bloom's taxonomy remains a hierarchical framework for organizing assessment objectives [22], yet evidence suggests that cognitive-process classifications cannot be reliably reduced to action verbs alone [30], and many assessment instruments still concentrate heavily on lower cognitive levels rather than spanning the full hierarchy [42].

This creates a structural risk for adaptive systems: when heterogeneous evidence is aggregated into a single performance representation, strong performance at one cognitive level may mask instability at another, allowing progression decisions that are psychometrically efficient but pedagogically incoherent. Existing rating-based adaptive systems, including Elo and Glicko variants, support dynamic learner estimation and efficient online updating, and Glicko-based approaches explicitly model confidence in the current rating [1], [58]. However, recent reviews of learner modeling and adaptive assessment still describe a landscape organized around model families, algorithms, and personalization mechanisms rather than architectures that formalize Bloom hierarchy itself as a non-compensatory progression constraint [6], [22]. Although prior work has shown that Elo-based learner modeling can be effectively applied in adaptive educational systems [39], these approaches generally remain single-dimensional and do not govern advancement through explicit uncertainty-bounded progression rules. Bloom-related work in this line has also treated cognitive taxonomy primarily as a way of characterizing task complexity rather than as a formal control structure for adaptive progression [41].

To address this gap, this study proposes the Hierarchical Bloom-Constrained Glicko (HBC-Glicko) architecture, in which learner ability is represented as a band-specific vector aligned with Bloom's cognitive structure, and advancement is controlled through anchor-derived readiness thresholds combined with confidence-bound constraints. In this way, progression depends not only on estimated ability but also on evidential stability at prerequisite levels. The contribution is conceptual and architectural rather than empirical: the study formalizes a measurement architecture and derives its principal theoretical properties, establishing a foundation for subsequent simulation and empirical investigation.

*Corresponding author.

More specifically, the present work pursues three objectives:

O1. To formalize an adaptive assessment architecture in which Bloom's revised taxonomy operates as a non-compensatory structural constraint on learner progression, rather than as a descriptive classification of task difficulty.

O2. To derive the principal theoretical properties that follow from combining band-specific learner representation, Glicko-based uncertainty modeling, and confidence-bound progression rules within a single measurement framework.

O3. To situate the proposed architecture relative to existing rating-based learner models and clarify the structural distinctions it introduces with respect to cognitive hierarchy, uncertainty quantification, and progression control.

The remainder of the study is organized as follows. Section II develops the theoretical rationale grounding the architecture. Section III presents the overall system architecture. Sections III-A through III-E formalize the learner model, within-band adaptation, anchor-based thresholds, hierarchical decision rules, and inactivity-driven uncertainty growth, collectively addressing the first objective. Section IV derives the principal architectural properties of the design (Objective 2). Section V situates HBC-Glicko relative to existing approaches and discusses its theoretical implications (Objective 3). Sections VI and VII address limitations and directions for future research.

II. Background and Theoretical Rationale

A. Adaptive Assessment and the Limits of Scalar Learner Models

Adaptive formative assessment systems personalize item delivery by combining online learner-state estimation with sequential item selection, commonly operationalized through IRT-based computerized adaptive testing, Elo-style rating updates, and Bayesian knowledge tracing. In computerized adaptive testing (CAT), learner performance is typically summarized by a single latent proficiency parameter θ under unidimensional item response theory [62], an approach whose feasibility continues to be demonstrated in large-scale digital applications [9].

However, this scalar abstraction assumes that performance can be adequately explained along a single proficiency continuum, an assumption that holds only when the measured construct is approximately unidimensional [65]. When tasks differ in cognitive process requirements, collapsing performance into one dimension produces compensatory interpretations in which strengths in one type of task mask weaknesses in another. Multidimensional adaptive frameworks address this limitation by estimating multiple latent traits, including approaches that represent uncertainty in multivariate ability estimation through confidence regions [31]. A similar limitation appears in rating-based learner models: educational adaptations of Elo efficiently update learner and item estimates after each interaction [58] but generally summarize learner state as a single scalar rating without encoding variation in cognitive demand [1].

Knowledge-component and skill-hierarchy models provide a further alternative to scalar learner models by representing learner performance in relation to discrete skills or prerequisite structures rather than as a single global proficiency estimate [12], [10], [38], [63]. Their main strength is that they can support adaptive sequencing by tracking evidence of mastery across specific components.

However, their dependency structure usually concerns concepts, procedures, or skills rather than Bloom-level cognitive-process demands. From a progression-control perspective, the key limitation is architectural: these models may estimate what a learner is likely to know or perform, but they do not define when the learner has demonstrated sufficiently stable competence at one cognitive level to justify advancement to the next. To our knowledge, existing IRT-, Elo-, and knowledge-component-based architectures do not treat Bloom's hierarchy as an internal non-compensatory constraint on the learner model itself. When cognitive levels appear, they function as item descriptors or reporting categories rather than as confidence-bounded rules governing progression. This limitation motivates learner representations that preserve distinctions in cognitive demand while retaining online updating and uncertainty quantification.

B. Revised Bloom Taxonomy and Hierarchical Dependency

Bloom's revised taxonomy (RBT) distinguishes a cognitive-process dimension (Remember, Understand, Apply, Analyze, Evaluate, and Create) and a knowledge-type dimension (Factual, Conceptual, Procedural, and Metacognitive), supporting constructive alignment between objectives, instruction, and assessment evidence [3], [28]. However, classification validity may weaken when cognitive level is inferred from surface cues such as verbs rather than from the actual mental operations required by the task [30]. Assessment instruments often fail to reflect the intended balance of cognitive demand: examinations may become dominated by lower-level recall [35], or higher-order demands may increase in ways that raise concerns about fairness [53].

Deeper understanding is better captured when assessment design includes integrative and evaluative demands [34]. This is especially consequential for rating-based learner models, where items are reduced to a single difficulty parameter that may obscure meaningful differences in cognitive demand.

For the present work, RBT is not treated as a rigid staircase but as a plausible dependency structure in which higher-order performance is generally more likely when foundational representations are sufficiently established. This interpretation is consistent with classic validation work suggesting cumulative relations across broad groupings rather than strictly between every adjacent pair [29], and with more recent evidence showing that higher-level objectives may depend on lower ones through indirect pathways [60]. From a cognitive perspective, this is plausible because more complex reasoning places greater demands on working memory and becomes more feasible when prior knowledge is organized into schemas that reduce cognitive load [50], [45].

C. Uncertainty-Aware Progression in Rating-Based Systems

Elo-style rating systems are widely used in adaptive educational systems because they update learner and item parameters online with modest computational cost [24], [39],

[36], [2], [40]. However, standard Elo represents learner state as a single point estimate without an explicit uncertainty parameter, which can produce unstable early estimates and, in adaptive settings that jointly update learners and items, rating variance that inflates over time [7]. These models also generally ignore the lag between successive interactions, treating adjacent observations as if no learning or forgetting occurred between them [12], [56], [10], [38].

These limitations motivate uncertainty-aware approaches that attach reliability information to ability estimates [21]. One strategy replaces the constant update factor with a decreasing uncertainty function that reduces rating sensitivity as interactions accumulate [2], [57]. The Glicko rating system formalizes this idea by introducing rating deviation (RD) as an explicit uncertainty parameter: high RD permits larger updates under limited evidence, inactivity, or cold start, whereas decreasing RD stabilizes the estimate and supports more conservative adaptive decisions [17].

However, the deeper limitation of Elo and Glicko variants in cognitively structured domains is not merely technical but representational. These systems originate from paired-comparison rating contexts in which the primary objective is to estimate relative performance or predict outcomes. When transferred to education, however, the rating system must support a different kind of decision: whether the learner is pedagogically ready to progress. This requires more than adding educational labels to a generic rating scale. It requires a learner model whose structure reflects educational constraints, including cognitive demand, prerequisite stability, and uncertainty-aware progression.

Recent extensions have introduced multivariate Elo and Glicko representations that estimate separate ratings for knowledge components [1], [2]. These approaches improve diagnostic resolution, but they retain a primarily estimation-oriented design: the multiple dimensions represent knowledge components whose ratings are updated separately, and in some cases cross-informed, rather than an ordered cognitive hierarchy in which advancement is governed by explicit readiness conditions. Thus, when hierarchy is present, it mainly supports parameter estimation rather than constraining progression decisions. This distinction is central to the present work: HBC-Glicko is not simply Bloom-aware estimation, but Bloom-constrained progression.

In scalar Elo and Glicko variants, learner state is encoded as a single rating, so the system has no internal mechanism to detect whether strong performance at one cognitive level is masking weakness at another. In a Bloom-structured formative setting, this can permit advancement on the basis of aggregated performance while foundational knowledge remains unstable, producing a progression decision that is psychometrically plausible but pedagogically incoherent.

D. Binary and Continuous Evidence in Adaptive Assessment

The representation of learner evidence should reflect the cognitive complexity of the task being assessed. For lower Bloom levels, particularly Remember and Understand, binary scoring (correct/incorrect) is computationally efficient and aligns with conventional objective formats. However, although multiple-choice questions can theoretically assess higher-order thinking, empirical work shows they often concentrate on lower-order demands in practice [42], [52], [48], [64], [33]. Higher Bloom levels require richer evidence because responses involve reasoning, justification, or production that cannot be reduced to binary judgments. Rubric-based scoring addresses this by assessing performance against explicit criteria and levels of quality [23], [8].

Rubric-based assessment poses a scalability challenge in formative contexts where consistent human scoring of large response volumes becomes impractical. Large language models offer a promising solution: recent studies suggest that LLMs can support rubric-based scoring at scale with acceptable reliability for formative use [11], [14], [16], [47]. These considerations support a mixed evidence design, with binary scoring for lower Bloom levels and rubric-based continuous evidence for higher-order cognitive performance.

Taken together, the preceding sections suggest that adaptive progression in hierarchical cognitive domains cannot be governed by scalar performance estimates alone. The system must account for hierarchical dependency, uncertainty in learner estimates, and differences in evidence type across task formats. These requirements motivate an explicit control logic that regulates advancement conservatively rather than treating progression as a simple consequence of observed success. The next section presents the system architecture and formal control design of the proposed HBC-Glicko system.

III. System Architecture Overview

To clarify how the proposed components interact, this section presents the overall system architecture of HBC-Glicko (Fig. 1). The architecture provides a high-level view of the information flow from the calibrated, Bloom-tagged item bank to item selection, scoring, rating updates, and progression decisions. This overview is intended to orient the reader before the formal definitions and control rules introduced in the following sections.

A. Operational Deployment Flow

To make the architecture more operationally explicit, Fig. 2 presents a practical deployment cycle for HBC-Glicko in an LMS-supported formative assessment context. The cycle begins with LLM-assisted item generation and Bloom-label suggestion, followed by instructor review and confirmation to ensure that each item is pedagogically appropriate and correctly assigned to a cognitive band and category. Validated items then enter a calibrated Bloom-banded item pool with anchor sets used for progression thresholds. During assessment, the system routes the learner within the current band, selects and scores items, and commits band-specific rating updates at the session level. Progression decisions are then made through confidence-bound rules: the learner is promoted only when readiness is sufficiently stable, whereas prerequisite reinforcement is triggered when current-band performance falls below the expected entry standard. This deployment view clarifies that HBC-Glicko functions not as an autonomous replacement for instructional judgment, but as a supervised adaptive layer that combines

LLM-supported item preparation, instructor validation, and uncertainty-aware progression control.

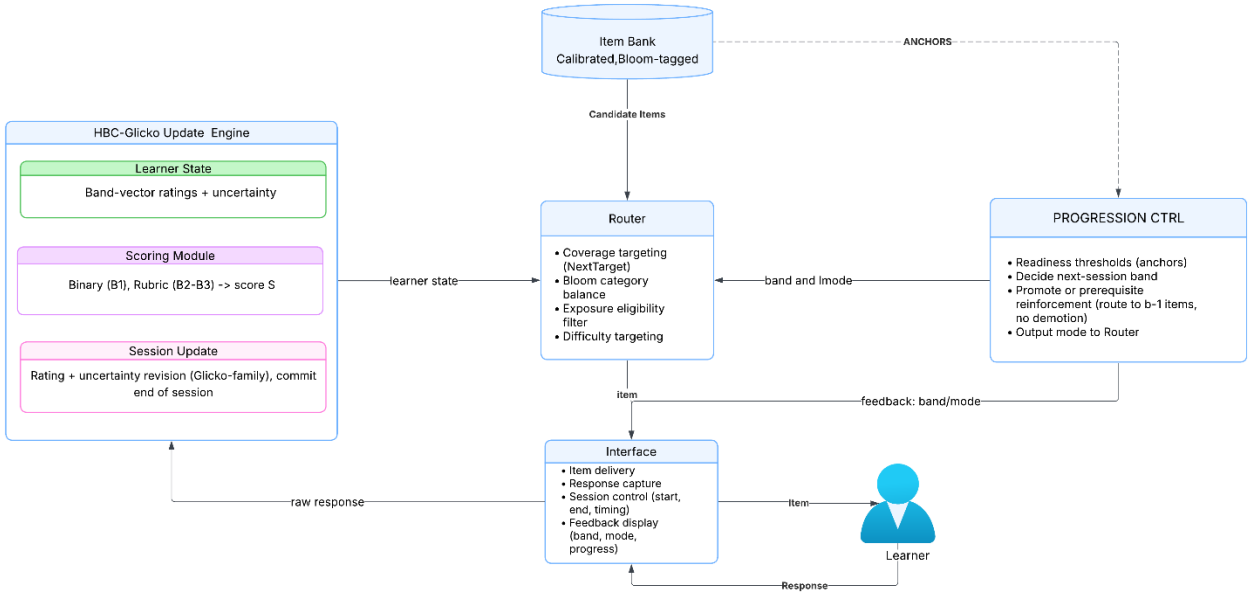


Fig. 1. Overview of the HBC-Glicko architecture.

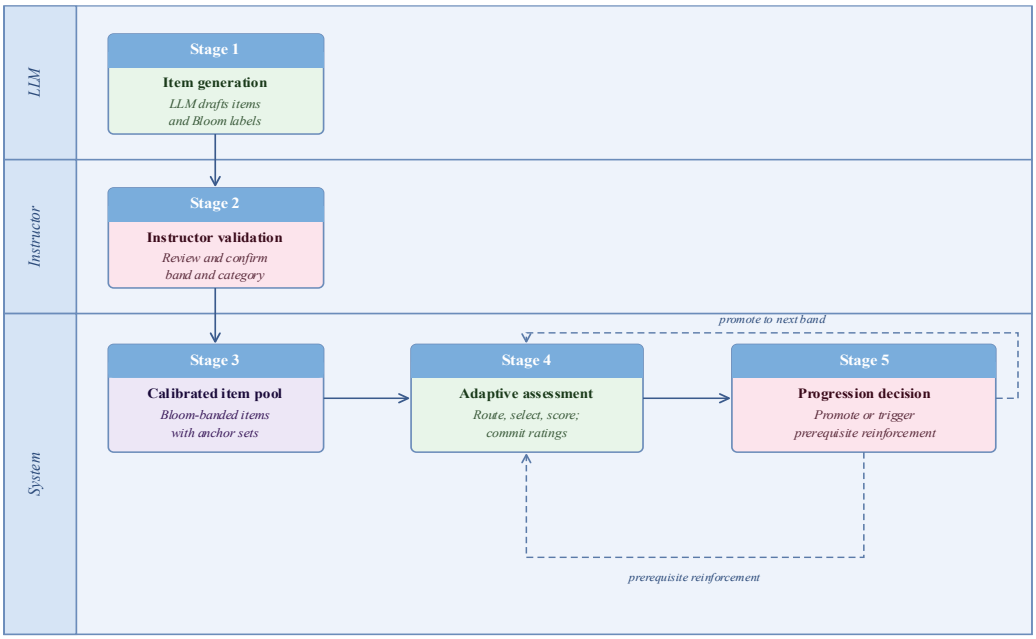


Fig. 2. HBC-Glicko deployment cycle. Dashed arrows indicate conditional system-level transitions after Stage 5.

The proposed system integrates five components, each developed in the following subsections: the formal learner model (Section III-A), within-band adaptation (Section III-B), anchor-based progression thresholds (Section III-C), hierarchical decision rules for promotion and prerequisite reinforcement (Section III-D), and inactivity-driven uncertainty growth (Section III-E).

1) Formal learner measurement model: This section specifies the formal representation of a learner within the proposed adaptive measurement architecture. The learner is defined as a structured latent state with band-specific performance parameters aligned with Bloom-type cognitive bands. The purpose is to define 1) the learner state variables, 2) their uncertainty representation, and 3) the structural assumptions that prevent compensatory masking across bands. This section provides definitions only. Update equations, thresholds, and decision rules are specified in subsequent sections.

For reference, a complete summary of the notation used throughout the following sections is provided in Appendix A (Table III).

2) Band set and indexing: Let n index learners, t index rating periods (sessions), and b ∈ {1, 2, 3} index cognitive bands. The bands represent increasing levels of cognitive complexity, following the revised Bloom taxonomy [3]. Although the revised taxonomy distinguishes two dimensions — cognitive processes and knowledge types — this system adopts the cognitive-process dimension only, consistent with common practice in empirical assessment research [59], [35], [5].

Low band: $\mathcal{K}$_Low = {Remember, Understand}, Mid band: $\mathcal{K}$_Mid = {Apply}, High band: $\mathcal{K}$_High = {Analyze, Evaluate}.

The conventional distinction in educational practice divides these categories into lower-order thinking skills (Remember, Understand, and Apply) and higher-order thinking skills (Analyze, Evaluate, and Create) [28]. While useful for curricular planning, this two-tier partition does not distinguish between retrieval, comprehension, and contextual transfer, which are diagnostically different in adaptive assessment. The present architecture groups the categories according to the type of cognitive operation involved. The Low-band combines Remember and Understand as representational processes in which the learner accesses or restates stored knowledge. The Mid-band isolates apply as the transition from accessing knowledge to using it under unfamiliar conditions. The High band combines Analyze and Evaluate as processes that require coordinating and judging multiple knowledge elements simultaneously, imposing high working memory demand [49]. This partition captures three qualitatively distinct modes of cognitive engagement while preserving sufficient evidence per band for stable rating estimation in short formative sessions.

Create is excluded because it introduces constraints that differ in kind from the other levels. Creative production tasks require the learner to generate an original artifact rather than respond to a structured prompt, which means that item difficulty cannot be stably pre-calibrated on the unified rating scale: the effective demand of a Create task depends on the scope and ambition of the learner's response, not solely on fixed item properties. Additionally, while LLM-assisted scoring is used in higher bands for evaluating reasoning quality against explicit rubric criteria, scoring creative output requires additional appraisal of novelty and integrative originality, dimensions for which automated scoring consistency remains less established [8]. Incorporating Create would therefore require modifications to both the item calibration model and the scoring framework, constituting a substantive extension left for future work.

Each item in the pool carries two fixed labels: a band (b) and a Bloom category within that band. These labels are assigned through a two-stage process. First, an LLM classifier suggests both the band and the category. Second, the instructor reviews and confirms the suggestion before the item enters the pool. Once confirmed, neither label is changed by the system at runtime.

This dual labeling serves two purposes. The band label supports the non-compensatory structure of the learner model. The within-band category label supports the within-band coverage mechanism.

3) Learner state variables: For each learner $\boldsymbol{n}$ at the rating period $\boldsymbol{t}$, the latent performance state is represented by $\mathbf{r}_{\boldsymbol{n,t}} = (\boldsymbol{r_{n,1,t}}, \boldsymbol{r_{n,2,t}}, \boldsymbol{r_{n,3,t}})$, where $\boldsymbol{r_{n,b,t}} \in \mathbb{R}$ denotes performance in band $\boldsymbol{b}$. Each estimate is associated with a rating deviation $\boldsymbol{RD_{n,b,t} > 0}$, which captures uncertainty. The learner measurement state is

$$\boldsymbol{\mathcal{S}_{n,t}} = \{(\boldsymbol{r}_{n,b,t}, \boldsymbol{RD}_{n,b,t}) : \boldsymbol{b} \in \boldsymbol{\mathcal{B}}\}.$$

Consistent with the theoretical rationale developed in Section II, learner state is modeled as a vector rather than as a single scalar proficiency parameter. This choice reflects the view that performance depends on prior knowledge organization and on the cognitive demands imposed by different tasks, rather than on a single context-free ability alone [49], [45]. High performance in a higher band, therefore, cannot substitute for insufficient performance in a lower band. A non-compensatory representation is retained to preserve pedagogical coherence in progression decisions and to prevent apparent success on more complex tasks from masking weakness in foundational knowledge.

4) Unified scale assumption: All band ratings are expressed on a common continuous metric scale. This enables coherent definition of item difficulty parameters, readiness thresholds, and uncertainty bounds on a single axis, while preserving band separation through the evidence structure.

Let $d_i \in \mathbb{R}$ denote the difficulty (location) parameter of item i on the same scale as $r_{n,b,t}$. The architecture assumes that expected performance is a monotonic function of the rating difference $r_{n,b,t} - d_i$. The explicit response function and update rules are provided in Section III-B. Here, the key assumption is scale compatibility: learner and item parameters share a metric that supports probabilistic interpretation.

B. Within-Band Adaptation

This section specifies the within-band probabilistic update mechanism. Each cognitive band operates as an independent rating process governed by a Glicko-style uncertainty-aware model [18], [19]. Interactions with items assigned to the band b update only the learner parameters $(r_{n,b,t} | RD_{n,b,t})$ associated with that band and the parameters of the interacted item.

1) Response model: Consider an item i in band b, with difficulty $d_{i,t} \in \mathbb{R}$ and rating deviation $RD_{i,t} > 0$. Let $r_{n,b,t}$ denote learner n's rating in the band b. Following the Glicko formulation, define.

$$g(RD_{i,t}) = \frac{1}{\sqrt{1 + \frac{3q^2 RD_{i,t}^2}{\pi^2}}}, q = \frac{\ln(10)}{400}$$

The expected score of the learner n on item i at time t is

$$E_{n,i,t} = \frac{1}{1 + 10^{-\frac{g(RD_{i,t})(r_{n,b,t} - d_{i,t})}{400}}}$$

Observed performance is represented on a common scale $S_{n,i,t} \in [0,1]$. This common scale is retained to preserve comparability across bands while allowing the observable response format to remain aligned with the cognitive demands of the task. In Band 1, where tasks primarily target retrieval and foundational understanding, responses are binary, with $S_{n,i,t} \in \{0,1\}$, so $E_{n,i,t}$ corresponds to the probability of a correct response. In Bands 2 and 3, where performance depends more on the quality of application, analysis, justification, or evaluation, rubric-based scores are normalized to [0, 1]. In those bands, $E_{n,i,t}$ is interpreted as the expected performance level on the same scale.

This mixed-evidence design is intended to ensure that the granularity of scored evidence remains congruent with the level of cognitive complexity being assessed. Lower-band performance can often be adequately captured through correctness, whereas higher-band performance requires attention to gradations in the quality of reasoning and knowledge use. The purpose of this design is therefore pedagogical as well as technical. By matching the scoring format to the nature of the targeted performance, the system can support formative interpretations that distinguish between foundational correctness and higher-order quality of reasoning, which is important for making progress intelligible to learners across increasing levels of task complexity [37], [54].

The prediction error is then defined as

$$\delta_{n,i,t} = S_{n,i,t} - E_{n,i,t}.$$

This quantity represents the deviation between observed and expected performance and serves as the input to the subsequent update step.

2) Learner rating update: Within band b, the learner rating is updated as

$$r_{n,b,t+1} = r_{n,b,t} + \omega_{n,b,t}\delta_{n,i,t}$$

where, $\boldsymbol{\omega_{n,b,t} > 0}$ denotes a Glicko-style adaptive weight. The update magnitude increases with learner uncertainty in the band $\boldsymbol{b}$, decreases with item uncertainty through $\boldsymbol{g(RD_{i,t})}$, and depends on the interaction-information term $\boldsymbol{E_{n,i,t}(1 - E_{n,i,t})}$, which reaches its maximum when the predicted outcome is most uncertain. Band updates are strictly independent, such that an interaction in band $\boldsymbol{b}$ affects only the learner state $\boldsymbol{(r_{n,b,t'} RD_{n,b,t})}$ for that band and the associated item parameters.

3) Within-band category routing and item selection: Within a session, the active band $b \in$ {Low, Mid, High} remains fixed. The routing mechanism operates only within the current band and determines, at each step t, which in-band Bloom category should be sampled next.

a) Phase 1: Target Category Determination: Phase 1 determines the next category to be sampled within the active band. To guide this decision, the instructor specifies a target sampling distribution over the categories in $\mathcal{K}_b$, denoted $\pi_b(k)$, such that $\pi_b(k) \geq 0$ and $\sum_{k \in \mathcal{K}_b} \pi_b(k) = 1$. By default, this distribution is uniform, although it may be adjusted to reflect pedagogical priorities. During the session, the system maintains, for each category $k \in \mathcal{K}_b$, a delivery counter $c_t(k)$ representing the number of scored items delivered up to step t in band b, with total in-band exposure given by $t_b = \sum_{k \in \mathcal{K}_b} c_t(k)$. The expected count for the category k under the target distribution is $\hat{c}_t(k) = t_b \pi_b(k)$, and the corresponding coverage deficit is defined as $\Delta_t(k) = \max(0, \hat{c}_t(k) - c_t(k))$. The next target category is then selected as the most under-sampled category, $k_t = \arg\max_{k \in \mathcal{K}_b} \Delta_t(k)$. In the Mid band, where $\mathcal{K}_{Mid} = \{Apply\}$, this step is trivial, since $k_t = Apply$ at every step. If multiple categories have the same deficit, priority is given to the category with the larger remaining item pool, and any further ties are resolved deterministically to ensure reproducibility.

b) Phase 2: Adaptive Item Selection Within the Target Category: Once the target category k_t has been determined, the router forms the candidate set $I_t^{cand} = \{ i : b(i) = b, k(i) = k_t \}$. Because k_t is selected in Phase 1, this set is already aligned with the remaining within-band category demand. To preserve measurement independence within a session, only the first scored attempt on an item is allowed to update ratings. Let H_s denote the set of items already scored in the session s. The eligible set is then $I_t^{elig} = \{ i \in I_t^{cand} : i \notin H_s \}$. If $I_t^{elig} = \emptyset$, secondary constraints may be relaxed in a deterministic manner while preserving the band constraint. This rule is motivated by a conservative cognitive rationale: repeated responses to the same item within a session may still draw on information that remains readily accessible in working memory, whereas a first scored attempt in a later session is more plausibly treated as a new retrieval event from long-term memory [4], [13]. This distinction is also consistent with retrieval-practice research showing that immediate repetition and later retrieval do not have the same implications for retention [43], [25]. Within the eligible set, the next item is selected using a probability-targeting principle. Let $E_{n,i,t}$ denote the expected performance of the learner n on item i at step t, and let $p^* \in (0,1)$ denote the target performance level. In Band 1, $E_{n,i,t}$ is the probability of a correct response under a binary outcome model. In Bands 2 and 3, where scores are rubric-based and normalized to [0,1], it represents the expected performance level. The selected item is therefore $i_t^* = \arg\min_{i \in I_t^{elig}} | E_{n,i,t} - p^* |$. If several items are equally close to p^*, the tie is broken using the information function $Info(i) = g(RD_{i,t})^2 E_{n,i,t}(1 - E_{n,i,t})$, and the item with the highest value is selected.

4) End-of-session rating commitment: Although provisional parameter adjustments may be computed during a session for routing purposes, official rating updates are committed only at the end of the session. All scored interactions from the session are aggregated, and the band-specific learner parameters $\boldsymbol{(r_{(n,b)}, RD_{(n,b)})}$ are updated from their pre-session values using the complete set of session evidence.

This separation serves both a statistical and a pedagogical purpose. Statistically, it avoids sequential double counting

within the same session and preserves coherence in parameter estimation. Pedagogically, it reduces the risk that transient within-session fluctuations are overinterpreted as stable change in learner state. By committing ratings only after the full session has been observed, the model treats learner performance as session-level evidence rather than as a sequence of isolated micro-updates, which is more consistent with the formative aim of basing progression decisions on sufficiently consolidated patterns of performance rather than on momentary variation.

C. Anchor-Based Progression Thresholds

1) Entry anchor sets and reference difficulty definition: In educational measurement, anchor items are used to preserve scale stability across forms and administrations by maintaining item and ability estimates on a shared metric [26], [55]. The present architecture adopts this principle to define stable progression thresholds across Bloom bands. For each transition from band $b$to band $b+1$, an entry anchor set $\mathcal{A}_{b+1} \subset \mathcal{I}_{b+1}$is defined as a fixed subset of calibrated items drawn exclusively from band $b+1$, where $\mathcal{I}_{b+1}$denotes the item pool associated with that band. Each anchor item $j \in \mathcal{A}_{b+1}$is characterized by a calibrated difficulty parameter $d_{b+1,j}$on the unified Glicko scale, so that anchor difficulties share the same latent metric as learner band-specific ratings. Because these anchor sets are predefined for progression purposes and are not dynamically altered by session-level adaptation, they serve as stable structural reference points for entry-level demand in the target band. The reference difficulty associated with entry into band $b+1$is defined as the mean anchor difficulty,

$$d_{b+1}^{ref} = \frac{1}{|\mathcal{A}_{b+1}|} \sum_{j \in \mathcal{A}_{b+1}} d_{b+1,j}$$

This quantity represents the typical calibrated difficulty of entry-level tasks in band $b+1$ and provides a fixed scale location against which readiness criteria are defined. In pedagogical terms, it operationalizes the level of demand that a learner should be prepared to meet before progression is warranted. To preserve the interpretability of progression thresholds over time, anchor item parameters are treated as fixed after calibration and excluded from routine item-parameter updating. In this way, anchor-based reference locations remain effectively stable on the unified rating scale, ensuring that subsequent readiness decisions retain a consistent measurement meaning.

2) Probability-defined readiness thresholds: Progression from band $b$to band $b+1$is defined in probabilistic terms. Let $p_{b\to b+1} \in (0,1)$denote the target expected success probability required for entry into band $b+1$. This parameter specifies the minimum expected performance required on representative entry-level tasks of the target band.

Readiness is defined relative to the reference difficulty d_{b+1}^{ref} established above. Let $r_{n,b}$ denote learner n 's ability estimate in band b. On the unified rating scale, the expected outcome of the learner $n$on an item of difficulty $d$is given by

$$E(r_{n,b}, d) = \frac{1}{1 + 10^{-\frac{r_{n,b}-d}{400}}}.$$

The readiness criterion requires expected performance at the target-band reference difficulty to equal the designated transition probability:

$$E\left(r_{n,b}, d_{b+1}^{ref}\right) = p_{b\to b+1}.$$

Solving for the rating value that satisfies this condition yields the progression threshold

$$T_{b\to b+1} = d_{b+1}^{ref} + 400 \log_{10}\left(\frac{p_{b\to b+1}}{1 - p_{b\to b+1}}\right).$$

Thus, $T_{b\to b+1}$ represents the minimum band- b rating required for a learner to achieve the target expected success probability on calibrated entry-level tasks from band $b+1$. Progression is therefore defined as a predictive readiness condition on the unified scale rather than as a retrospective summary of past performance. This predictive interpretation is consistent with cognitivist accounts of learning in which successful future performance depends on the interaction between prior knowledge organization and the cognitive demands of upcoming tasks, rather than on past score accumulation alone [49], [45].

In contrast to standard CAT item selection, which typically targets items near the current ability estimate to maximize measurement information, progression thresholds in the HBC system are designed to express readiness for cognitively more demanding work. Accordingly, transition probabilities are specified above the chance level and increase across bands:

$$0.50 < p_{Low\to Mid} < p_{Mid\to High} < 1.$$

This reflects the design assumption that progression to higher cognitive bands should require stronger evidence of readiness. Pedagogically, this means that advancement into more demanding bands is treated as a guarded instructional decision, since the cost of premature progression increases as task complexity and knowledge coordination demands rise [49]. The specific values of $p_{b\to b+1}$are treated as design parameters to be calibrated empirically rather than fixed a priori.

D. Confidence-Bound Decision Rules for Band Progression

1) Rating uncertainty and confidence bounds: To incorporate estimation uncertainty into band progression, lower and upper uncertainty bounds are defined as

$$LCB_{n,b} = r_{n,b} - z_b RD_{n,b}, \quad UCB_{n,b} = r_{n,b} + z_b RD_{n,b}$$

where, $z_b > 0$ controls the width of the interval. The lower bound provides a conservative estimate of learner ability in band b, whereas the upper bound provides a corresponding optimistic estimate. As $RD_{n,b}$ increases, the interval widens, which discourages progression under high uncertainty. As $RD_{n,b}$ decreases, the interval narrows, allowing readiness to be evaluated more decisively. This formulation makes readiness conditional not only on the estimated level, but also on the stability of the evidence supporting that estimate.

The multiplier $\boldsymbol{z_b}$ is set to 1.65, corresponding to a one-sided 95% confidence level under the distributional assumptions of

the Glicko framework [18]. Promotion therefore requires 95% confidence that the learner's true ability exceeds the readiness threshold, whereas reinforcement is triggered only when even a 95th percentile estimate falls below the entry standard. This moderately conservative threshold is consistent with the formative orientation of the architecture. A stricter bound, such as $z_b = 2.00$, would systematically delay promotion for learners who have already demonstrated sufficient mastery, moving them away from their zone of proximal development without proportionally reducing misclassification risk. The architecture can tolerate a less restrictive threshold because prerequisite reinforcement [Section III-D(3)] provides a built-in safeguard. If promotion proves premature, the system can detect instability and redirect the learner to lower band items. In summative contexts, where no such remediation loop exists, a stricter value would be more appropriate. More generally, z_b is treated as a tunable design parameter whose optimal value should be calibrated empirically for each deployment context. Although a uniform multiplier is adopted here for simplicity, z_b could in principle be differentiated across Bloom bands. A stricter bound at higher cognitive levels would reflect the greater pedagogical cost of premature advancement into more demanding tasks, with band-specific values determined through pilot calibration against observed promotion rates and downstream performance outcomes. The relationship between RD and the progression threshold is structural. Because $LCB_{n,b} = r_{n,b} - z_b \cdot RD_{n,b}$, higher uncertainty directly lowers the conservative readiness estimate and makes the promotion condition $LCB_{n,b} \geq T_{b\to b+1}$ harder to satisfy. A learner with a high point estimate but high $RD_{n,b}$ is therefore treated as less ready than a learner with the same point estimate and lower uncertainty. This is pedagogically intentional: promotion requires not only high estimated performance but also stable evidence within the prerequisite band. Formally, $T_{b\to b+1}$ is not a free parameter, but a derived quantity computed from d_{b+1}^{ref} and $p_{b\to b+1}$ according to the anchor-based threshold construction in Section III-C.

2) Promotion rule: conservative and coverage-constrained: Let $T_{b\to b+1}$ denote the readiness threshold for progression from band b to band $b+1$, and let $LCB_{n,b}$ denote the lower confidence bound of learner n's rating in band b. Promotion is granted only when the learner satisfies the conservative readiness condition.

$$\text{Promote } b \to b+1 \iff LCB_{n,b} \geq T_{b\to b+1}$$

To ensure that progression is not driven by performance in only a subset of within-band categories, promotion is additionally subject to a minimum evidence requirement over the category set $\mathcal{K}_b$. Let $N_{n,b}(k)$ denote the number of distinct scored item administrations obtained by learner n in category $k \in \mathcal{K}_b$, and let m_b denote the minimum required evidence per category. The full promotion rule is therefore

$$\begin{aligned}\text{Promote } b \to b+1 \\ \iff \left(LCB_{n,b} \geq T_{b\to b+1}\right) \\ \wedge \left(\forall k \in \mathcal{K}_b : N_{n,b}(k) \geq m_b\right)\end{aligned}$$

This rule makes progression both conservative and non-compensatory. The confidence-bound condition reduces the risk of premature advancement under high uncertainty, while the coverage condition ensures that promotion reflects sufficient evidence across all categories within the band rather than strong performance in only a limited subset.

3) Prerequisite reinforcement as a stability safeguard: Although promotion is irreversible, subsequent responses may indicate that the learner is not performing stably at the level expected for the current band. To detect such cases conservatively, the system uses the upper confidence bound of the band-specific rating. Prerequisite reinforcement is triggered when

$$UCB_{n,b} < T_{b-1\to b}$$

where, $T_{b-1\to b}$ denotes the entry threshold for band b. This condition indicates that, even under an optimistic estimate, the learner does not meet the readiness standard for the current band. When this occurs, the system temporarily prioritizes items from band $b-1$ to reinforce prerequisite knowledge. However, the learner's structural band membership remains unchanged, and no demotion is performed. Because $UCB_{n,b}$ increases with $RD_{n,b}$, reinforcement is less likely to be triggered when uncertainty is high, which reduces sensitivity to sparse or noisy evidence.

Together, the promotion and reinforcement rules define three decision zones. Promotion occurs when $LCB_{n,b} \geq T_{b\to b+1}$, subject to the coverage condition. Prerequisite reinforcement occurs when $UCB_{n,b} < T_{b-1\to b}$. Otherwise, the learner remains in band b without structural change.

E. Inactivity-Driven Uncertainty Growth

To reflect increasing uncertainty when no recent evidence is observed, rating deviation is allowed to grow during periods of inactivity. For learner n in band b, let $t_{n,b}$ denote the number of completed sessions since the learner last produced at least one scored attempt in that band. The effective pre-update uncertainty is defined as:

$$RD_{n,b}^{pre*} = \min\left(RD_{\max}, \sqrt{\left(RD_{n,b}^{pre}\right)^2 + c_b^2 t_{n,b}}\right)$$

where, $c_b > 0$ controls the rate of uncertainty growth under inactivity and $RD_{\max}$ is the maximum admissible rating deviation. This formulation allows uncertainty to increase as previously observed evidence becomes less informative over time, while preventing unbounded growth.

Because Band 1 is oriented toward retention-sensitive recall, the informativeness of prior evidence may decay more rapidly there than in higher cognitive bands [15], [46], [44], [27]. Accordingly, the inactivity-growth parameters are constrained such that

$$c_1 \geq c_2 \geq c_3,$$

so that uncertainty can grow at least as quickly in Band 1 as in Bands 2 and 3. If no new evidence is observed in band b, the band rating remains unchanged, and only the uncertainty term is updated through the inactivity-growth rule.

IV. Architectural Properties of HBC-Glicko

Table I summarizes the main theoretical properties that follow from the HBC-Glicko design. These properties are structural consequences of the learner-state representation, threshold formulation, confidence-bound decision rules, and inactivity update mechanism. They clarify what the architecture guarantees by construction, without implying empirical validation.

TABLE I. Main Theoretical Properties of the HBC-Glicko Architecture

Property	Architectural basis	Theoretical implication
Non-compensatory progression	Learner state is represented as a band-specific vector, and promotion is evaluated separately within each prerequisite band.	Strong performance in one band cannot compensate for insufficient evidence in another band.
Conservative promotion under uncertainty	Promotion depends on the lower confidence bound rather than the point estimate alone.	For the same rating estimate, greater uncertainty reduces effective readiness and makes promotion less likely.
Criterion-referenced progression	Promotion thresholds are defined by fixed anchor difficulties and target transition probabilities.	Readiness is judged against a common formal standard rather than a session-specific rule.
Asymmetric decision rules	Promotion uses the lower confidence bound, whereas prerequisite reinforcement uses the upper confidence bound.	The system is conservative at both boundaries: advancement requires sufficiently strong evidence, while remediation requires sufficiently weak evidence.
Inactivity-dependent uncertainty	Inactivity increases rating deviation without directly changing the stored rating.	After long inactivity, progression decisions become more cautious until new evidence is collected.

V. Discussion

The central contribution of HBC-Glicko is to reposition Bloom's revised taxonomy from a descriptive framework to a control principle within adaptive assessment. In many adaptive systems, the cognitive level is used to classify items or structure reporting, but it does not directly govern when a learner may advance. In the present architecture, Bloom-structured bands are embedded in the learner model itself, and progression depends on whether the learner has accumulated sufficiently strong and sufficiently stable evidence within the prerequisite band. Scalar learner models absorb performance across cognitively different tasks into a single estimate, allowing strength in one area to mask instability in another.

HBC-Glicko prevents this by representing learner state as a non-compensatory band vector and evaluating readiness within bands rather than across them [1], [62].

Table II situates HBC-Glicko relative to selected adaptive learner modeling approaches. None of the compared systems combines multidimensional Bloom-structured representation, explicit uncertainty quantification, and hierarchy used as a non-compensatory control constraint on progression within a single architecture.

TABLE II. Comparative Positioning of HBC-Glicko Relative to Selected Adaptive Learner Modeling Approaches

Feature	Standard Elo	Glicko	Multivariate Elo	Multivariate Glicko	CAT / UIRT	Skill-hierarchy Elo	Bloom-aware Elo	HBC-Glicko
	[39]	*[17], [18]*	*[2]*	*[1]*	*[62]*	*[63]*	*[20]*	*(proposed)*
Learner model type	Scalar	Scalar	Vector (KC)	Vector (KC)	Scalar	Multi-level (global + per-skill)	Vector (Bloom)	Vector (Bloom bands)
Explicit uncertainty parameter	No	Yes (RD)	Not explicit	Yes (RD per KC)	Yes (SEM)	No	No	Yes (RD per band)
Hierarchy represented	No	No	Skill components	Skill components	No	Skill hierarchy	Bloom levels	Bloom bands (ordered)
Hierarchy as control constraint	No	No	No	No	No	Estimation only	Estimation only	Yes
Non-compensatory progression	No	No	No	No	No	No	No	Yes
Anchor-based transition thresholds	No	No	No	No	No	No	No	Yes
Confidence-bounded progression rule	No	No	No	No	No	No	No	Yes
Explicit prerequisite reinforcement	No	No	No	No	No	No	No	Yes
Evidence differentiated by cognitive band	No	No	No	No	No	No	No	Yes

Note. KC = knowledge component. RD = rating deviation. SEM = standard error of measurement. UIRT = unidimensional item response theory. "Estimation only" indicates that the hierarchy informs parameter estimation or cross-level propagation but does not function as a constraint on learner progression. "Not explicit" indicates that the feature may be partially present or achievable through indirect mechanisms but is not formalized as a distinct architectural component. "Indirect (SEM)" reflects that CAT quantifies estimation precision through the standard error of measurement, which serves a different function from an explicit uncertainty parameter governing progression decisions.

A second contribution is that uncertainty is given a genuinely pedagogical function. In conventional Glicko, rating deviation determines update magnitude. In HBC-Glicko, uncertainty also determines what the system should do next: promotion depends on a lower-confidence interpretation of readiness, whereas prerequisite reinforcement is triggered only when performance remains weak even under an optimistic interpretation. This asymmetry makes the architecture deliberately conservative at transitions between cognitive bands, where instructional errors are most costly. Uncertainty does not merely slow or accelerate adaptation but directly constrains progression, extending its role from parameter updating to pedagogical decision-making.

To our knowledge, Hamzah and Sosnovsky (2023) are the only published work to directly combine a rating-based learner model with Bloom's revised taxonomy, and no prior work has combined a Glicko-based system with Bloom's taxonomy as a non-compensatory progression constraint. Yudelson et al. (2019) demonstrate that Elo can exploit skill hierarchies for few-shot estimation, and Abdi et al. [1] extend Glicko to multivariate knowledge components, but neither structure cognitive-process levels as a non-compensatory constraint on the learner model. In Hamzah and Sosnovsky (2023), Bloom levels organize learner modeling through cross-level propagation, so evidence at one level can influence estimates at another. In HBC-Glicko, bands are treated as non-compensatory decision spaces, and readiness for band b + 1 must be established from within band b itself. The distinction is therefore between Bloom-aware estimation and Bloom-constrained progression: the former improves the structure of the learner model; the latter governs advancement. For formative assessment, where sequencing and readiness are central, this is the more consequential step.

HBC-Glicko is not novel simply because it is multidimensional, since multivariate rating models already exist, nor because it incorporates uncertainty, since that is a defining feature of Glicko-based systems [1]. Its contribution lies in combining multidimensionality and uncertainty with a cognitively ordered progression rule. The dimensions are not parallel knowledge components estimated for predictive accuracy but ordered bands that participate directly in advancement logic, making the architecture more interpretable pedagogically than a generic multivariate model.

The contrast with computerized adaptive testing is equally important. CAT selects items that maximize information at the current ability estimate and achieves acceptable precision with few observations [62]. HBC-Glicko asks whether the learner has demonstrated enough stable evidence at the current cognitive band to justify movement to the next. This makes the architecture less efficient psychometrically but more coherent in formative contexts where progression should respect hierarchical cognitive dependency. HBC-Glicko should not be positioned as a replacement for CAT but as an alternative design philosophy for settings in which instructional ordering matters as much as measurement efficiency.

These advantages remain theoretical. The practical value of the architecture will depend on the quality of Bloom-level classification, the stability of anchor calibration, and the appropriateness of parameter choices governing thresholds, confidence penalties, and inactivity growth. The architecture nevertheless provides a formally specified and internally coherent foundation for future empirical work, with progression rules whose theoretical properties have been clarified in Section IV and whose design rationale is grounded in both measurement theory and cognitive theory. The value of HBC-Glicko will ultimately depend on how well those theoretical advantages translate into better progression decisions in practice.

Parameter Sensitivity and Calibration of Decision Behavior.

The proposed architecture makes progression decisions explicit by formalizing them through anchor-derived thresholds, confidence bounds, category-coverage requirements, and inactivity-driven uncertainty growth. This explicitness introduces parameter sensitivity, but it also constitutes an important strength of the model: the strictness of progression is no longer implicit, hidden in an opaque adaptive algorithm, or determined only by aggregated performance scores. Instead, it is represented through interpretable control parameters that can be inspected, adjusted, and empirically calibrated. The transition probability $p_{b\to b+1}$, the confidence multiplier z_b, the minimum evidence requirement m_b, and the inactivity-growth parameter c_b collectively determine the balance between adaptive responsiveness and instructional caution.

Misconfiguration of these parameters may shift the system toward predictable forms of decision error. If thresholds or confidence bounds are too conservative, HBC-Glicko may delay promotion for learners who are already capable of engaging with higher-order tasks. If they are too permissive, the system may allow premature progression before prerequisite competence is sufficiently stable. Similarly, an overly strict category-coverage requirement may produce unnecessary repetition, whereas an overly weak requirement may permit advancement from narrow evidence. Inactivity-growth parameters also require careful calibration: excessive growth may discount prior evidence too rapidly, while insufficient growth may cause the system to overtrust outdated performance. These risks do not undermine the architecture; rather, they indicate that HBC-Glicko should be treated as a tunable decision framework whose operating characteristics must be calibrated to the instructional context.

Future simulation and empirical work should therefore examine how alternative parameter settings affect promotion accuracy, reinforcement frequency, time-to-progression, learner workload, and downstream performance. Such analyses would make it possible to identify robust operating ranges and determine whether parameters should be band-specific. For example, stricter confidence bounds may be appropriate in

higher cognitive bands, where premature progression has greater pedagogical cost, whereas lower bands may require more responsiveness to avoid unnecessary repetition. In this sense, parameter sensitivity is not merely a limitation but a necessary design feature of an uncertainty-aware formative assessment architecture.

VI. Limitations

The most immediate limitation of this study is that HBC-Glicko remains entirely conceptual. No simulation studies, pilot deployments, or empirical datasets are presented to support the theoretical propositions outlined in Section IV. As a result, the practical advantages of the architecture cannot yet be confirmed beyond theoretical reasoning, and empirical testing remains a necessary next step before any performance claims can be made.

A second concern relates to the quality of Bloom-level classification. The architecture assumes that items are reliably tagged with the correct band and category labels. However, prior research has shown that cognitive-level classification is prone to error when inferred from surface features rather than the actual mental operations a task requires. Misclassified items could compromise the integrity of the band structure and produce progression decisions that are formally correct but educationally misleading.

A further limitation concerns parameter calibration. The architecture introduces several design parameters, including transition probabilities, confidence-bound multipliers, inactivity-growth rates, and minimum evidence requirements per category. Although these parameters make progression decisions explicit and adjustable, their optimal values cannot be established theoretically alone. Without simulation-based or empirical calibration, inappropriate parameter settings may make the system either overly conservative, delaying justified advancement, or overly permissive, allowing premature progression. Future work should therefore examine parameter sensitivity systematically and identify robust operating ranges across Bloom bands, domains, and learner populations.

Finally, the current architecture covers only three cognitive bands and does not incorporate the Create level or the knowledge-type dimension of Bloom's revised taxonomy. This restricts its applicability in learning contexts where generative or metacognitive performance is central. Extending the architecture to these dimensions represents a natural direction for future theoretical development.

VII. Directions for Future Research

Several research directions follow directly from the architecture proposed in this study. The most immediate priority is simulation-based validation. Agent-based simulations in which synthetic learners with known band-specific ability profiles interact with HBC-Glicko would make it possible to evaluate promotion accuracy, sensitivity to item mislabeling, convergence of band-specific ratings, and the effects of alternative parameter settings under controlled conditions. Such simulations would provide a necessary first test of the decision rules before deployment in live educational settings.

A second priority is empirical validation in authentic formative assessment contexts. Deploying HBC-Glicko alongside a conventional adaptive system would allow comparison of learning outcomes, progression patterns, and instructor judgments of diagnostic usefulness. It would also generate the real-world response data needed to calibrate anchor difficulties and transition probabilities.

A third direction concerns scoring validity in higher bands. Because the architecture relies on rubric-based evidence for more complex cognitive performance, and because such scoring may be partially supported by LLMs, future work should examine agreement between human and LLM-based ratings across Bloom levels and assess how any systematic divergence affects rating stability and progression decisions. This question is especially important given the mixed evidence design introduced in Section II-D.

Finally, the architecture could be extended in two theoretically important directions: inclusion of the Create level and incorporation of knowledge-type distinctions within bands. Both extensions could improve representational fidelity, but they would also increase model complexity and calibration demands.

VIII. Conclusion

This study introduced HBC-Glicko, a measurement architecture for formative adaptive assessment that embeds Bloom's revised taxonomy as a non-compensatory structural constraint within a Glicko-based learner model. The architecture formalizes learner state as a band-specific vector with explicit uncertainty, regulates progression through anchor-based readiness thresholds and confidence-bound decision rules, and models evidential decay under inactivity.

Three objectives were addressed. The study formalized cognitive hierarchy as a control principle governing adaptive progression rather than a descriptive classification of tasks. It derived the principal theoretical properties that follow from this design by construction. And it situates HBC-Glicko relative to existing rating-based approaches, clarifying that its contribution lies in combining multidimensional representation, uncertainty quantification, and non-compensatory progression within a single cognitively ordered architecture.

The contribution remains intentionally conceptual. Its practical value will depend on the reliability of item classification, the stability of anchor calibration, and the appropriateness of parameter configuration — conditions that can only be assessed through simulation and empirical deployment. These constitute the immediate next steps for this research program.

Conflict of Interest

The authors declare that there are no conflicts of interest related to this research.

AUTHORS' CONTRIBUTIONS

W.E.F. conceptualized the study, developed the HBC-Glicko architecture, formalized the theoretical model, and wrote the manuscript.

M.E. supervised the research, contributed to the conceptual refinement of the work, and provided critical revisions.

All authors reviewed and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENT

The authors extend their heartfelt gratitude to the National Center for Scientific and Technical Research (CNRST) for crucial support through the "PhD-Associate Scholarship – PASS" program. This support played a pivotal role in advancing the research and ensuring the successful completion of this study.

REFERENCES

[1] S. Abdi, H. Khosravi, and S. Sadiq, "Modelling learners in adaptive educational systems: A multivariate Glicko-based approach," in Proc. 11th Int. Learning Analytics and Knowledge Conf. (LAK '21), 2021, pp. 496–502. doi: 10.1145/3448139.3448189.

[2] S. Abdi et al., "A multivariate Elo-based learner model for adaptive educational systems," in Proc. 12th Int. Conf. Educational Data Mining (EDM 2019), 2019, pp. 228–233.

[3] L. W. Anderson and D. R. Krathwohl, Eds., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York, NY, USA: Longman, 2001.

[4] A. Baddeley, "Working memory: Theories, models, and controversies," Annu. Rev. Psychol., vol. 63, pp. 1–29, 2012. doi: 10.1146/annurev-psych-120710-100422.

[5] K. Banujan et al., "Revolutionising educational assessment: Automated question classification using Bloom's taxonomy and deep learning techniques," Int. J. Educ. Dev. Using Inf. Commun. Technol., vol. 19, no. 3, pp. 259–278, 2023.

[6] F. Böck et al., "Learner models: Design, components, structure, and modelling," User Model. User-Adapted Interact., 2025. doi: 10.1007/s11257-025-09434-4.

[7] M. Bolsinova, B. Gergely, and M. J. S. Brinkhuis, "Keeping Elo alive: Evaluating and improving measurement properties of learning systems based on Elo ratings," Brit. J. Math. Stat. Psychol., 2025. doi: 10.1111/bmsp.12395.

[8] S. M. Brookhart, "Appropriate criteria: Key to effective rubrics," Frontiers Educ., vol. 3, Art. 22, 2018. doi: 10.3389/feduc.2018.00022.

[9] W. Capon et al., "Validating an adaptive digital assessment of youth mental health needs: A cross-sectional study," npj Digit. Med., 2026. doi: 10.1038/s41746-026-02374-2.

[10] H. Cen, K. R. Koedinger, and B. Junker, "Learning factors analysis: A general method for cognitive model evaluation and improvement," in Intelligent Tutoring Systems (Lecture Notes Comput. Sci., vol. 4053), M. Ikeda, K. D. Ashley, and T. W. Chan, Eds. Berlin, Germany: Springer, 2006, pp. 164–175. doi: 10.1007/11774303_17.

[11] L.-H. Chang and F. Ginter, "Automatic short answer grading for Finnish with ChatGPT," Proc. AAAI Conf. Artif. Intell., vol. 38, no. 21, pp. 23173–23181, 2024. doi: 10.1609/aaai.v38i21.30363.

[12] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," User Model. User-Adapted Interact., vol. 4, no. 4, pp. 253–278, 1994. doi: 10.1007/BF01099821.

[13] N. Cowan, "Working memory underpins cognitive development, learning, and education," Educ. Psychol. Rev., vol. 26, no. 2, pp. 197–223, 2014. doi: 10.1007/s10648-013-9246-y.

[14] A. Divya, V. Haridas, and J. Narayanan, "Automation of short answer grading techniques: Comparative study using deep learning techniques," in Proc. 5th Int. Conf. Electrical, Computer and Communication Technologies (ICECCT), 2023, pp. 1–7.

[15] H. Ebbinghaus, Memory: A Contribution to Experimental Psychology, H. A. Ruger and C. E. Bussenius, Trans. New York, NY, USA: Columbia Univ. Press, 1913 (Original work published 1885).

[16] O. Fagbohun et al., "Beyond traditional assessment: Exploring the impact of large language models on grading practices," J. Artif. Intell. Mach. Learn. Data Sci., vol. 2, no. 1, pp. 1–8, 2024. doi: 10.51219/JAIMLD/oluwole-fagbohun/19.

[17] M. E. Glickman, "The Glicko system," Tech. Rep., Harvard Univ., Cambridge, MA, USA, 1995.

[18] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," J. Roy. Stat. Soc. C, vol. 48, no. 3, pp. 377–394, 1999. doi: 10.1111/1467-9876.00159.

[19] M. E. Glickman, "Dynamic paired comparison models with stochastic variances," J. Appl. Stat., vol. 28, no. 6, pp. 673–689, 2001. doi: 10.1080/02664760120059219.

[20] A. Hamzah and S. Sosnovsky, "Modelling student knowledge in blended learning," in Adj. Proc. 31st ACM Conf. User Modeling, Adaptation and Personalization (UMAP '23), 2023, pp. 76–80. doi: 10.1145/3563359.3597412.

[21] A. D. Hofman et al., "Tracking with (un)certainty," J. Intell., vol. 8, no. 1, Art. 10, 2020. doi: 10.3390/jintelligence8010010.

[22] A. Ihichr et al., "A systematic review on assessment in adaptive learning: Theories, algorithms and techniques," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 7, 2024. doi: 10.14569/IJACSA.2024.0150785.

[23] A. Jönsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," Educ. Res. Rev., vol. 2, no. 2, pp. 130–144, 2007. doi: 10.1016/j.edurev.2007.05.002.

[24] E. N. Kandemir et al., "Adaptation of the multi-concept multivariate Elo rating system to medical students' training data," in Proc. 14th Learning Analytics and Knowledge Conf. (LAK '24), 2024. doi: 10.1145/3636555.3636858.

[25] J. D. Karpicke and H. L. Roediger, "Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention," J. Exp. Psychol. Learn. Mem. Cogn., vol. 33, no. 4, pp. 704–719, 2007. doi: 10.1037/0278-7393.33.4.704.

[26] M. J. Kolen and R. L. Brennan, Test Equating, Scaling, and Linking: Methods and Practices, 3rd ed. New York, NY, USA: Springer, 2014. doi: 10.1007/978-1-4939-0317-7.

[27] N. Kornell, V. C. Rabelo, and A. T. Klingebiel, "Forgetting curves: New data and a new model," J. Mem. Lang., vol. 125, Art. 104303, 2022. doi: 10.1016/j.jml.2022.104303.

[28] D. R. Krathwohl, "A revision of Bloom's taxonomy: An overview," Theory Into Practice, vol. 41, no. 4, pp. 212–218, 2002. doi: 10.1207/s15430421tip4104_2.

[29] R. P. Kropp, H. W. Stoker, and W. L. Bashaw, "The validation of the taxonomy of educational objectives," J. Exp. Educ., vol. 34, no. 3, pp. 69–76, 1966. doi: 10.1080/00220973.1966.11010940.

[30] T. M. Larsen et al., "Probing internal assumptions of the revised Bloom's taxonomy," CBE—Life Sci. Educ., vol. 21, no. 4, Art. ar66, 2022. doi: 10.1187/cbe.20-08-0170.

[31] J. Li et al., "A new multidimensional computerized testing approach: On-the-fly assembled multistage adaptive testing based on multidimensional item response theory," Mathematics, vol. 13, no. 4, Art. 594, 2025. doi: 10.3390/math13040594.

[32] M. Liu et al., "Investigating the effect of an adaptive learning intervention on students' learning," Educ. Technol. Res. Dev., vol. 65, pp. 1605–1625, 2017.

[33] Q. Liu et al., "Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers," Innov. Educ. Teach. Int., vol. 61, no. 4, pp. 802–814, 2024. doi: 10.1080/14703297.2023.2222715.

[34] R. Lourdusamy, P. Magendiran, and C. M. Fonceca, "Analysis of cognitive levels of questions with Bloom's taxonomy: A case study," Int. J. Softw. Innov., vol. 10, no. 1, pp. 1–22, 2022. doi: 10.4018/IJSI.297922.

[35] T. Muhayimana, L. Kwizera, and M. R. Nyirahabimana, "Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools," Curriculum Perspectives, vol. 42, no. 1, pp. 51–63, 2022. doi: 10.1007/s41297-021-00156-2.

[36] M. Pankiewicz and M. Bator, "Elo rating algorithm for the purpose of measuring task difficulty in online learning environments," e-mentor, vol. 5, no. 82, pp. 43–51, 2019. doi: 10.15219/em82.1444.

[37] D. Parmigiani et al., "Formative assessment in higher education: An exploratory study within programs for professionals in education," Frontiers Educ., vol. 9, Art. 1366215, 2024. doi: 10.3389/feduc.2024.1366215.

[38] P. I. Pavlik Jr., H. Cen, and K. R. Koedinger, "Performance factors analysis: A new alternative to knowledge tracing," in Artificial Intelligence in Education, V. Dimitrova, R. Mizoguchi, B. du Boulay, and A. Graesser, Eds. Amsterdam, Netherlands: IOS Press, 2009, pp. 531–538.

[39] R. Pelánek, "Applications of the Elo rating system in adaptive educational systems," Comput. Educ., vol. 98, pp. 169–179, 2016. doi: 10.1016/j.compedu.2016.03.017.

[40] R. Pelánek et al., "Elo-based learner modeling for the adaptive practice of facts," User Model. User-Adapted Interact., vol. 27, no. 1, pp. 89–118, 2017. doi: 10.1007/s11257-016-9185-7.

[41] R. Pelánek, S. Silovská, and J. Řihák, "Complexity and difficulty of items in learning systems," Int. J. Artif. Intell. Educ., vol. 32, no. 4, pp. 1003–1048, 2022. doi: 10.1007/s40593-021-00252-4.

[42] H. Ravand et al., "Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's taxonomy," Frontiers Educ., vol. 10, Art. 1644811, 2025. doi: 10.3389/feduc.2025.1644811.

[43] H. L. Roediger and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," Psychol. Sci., vol. 17, no. 3, pp. 249–255, 2006. doi: 10.1111/j.1467-9280.2006.01693.x.

[44] R. Sacripante et al., "Forgetting rates of gist and peripheral episodic details in prose recall," Memory, vol. 31, no. 3, pp. 328–343, 2023. doi: 10.1080/09658211.2022.2052521.

[45] M. Schneider and B. A. Simonsmeier, "How does prior knowledge affect learning? A review of 16 mechanisms and a framework for future research," Learn. Individ. Differ., vol. 122, Art. 102744, 2025. doi: 10.1016/j.lindif.2025.102744.

[46] F. Sense et al., "An individual's rate of forgetting is stable over time but differs across materials," Topics Cogn. Sci., vol. 8, no. 1, pp. 305–321, 2016. doi: 10.1111/tops.12183.

[47] R. Sreedhar et al., "Comparing scoring consistency of large language models with faculty for formative assessments in medical education," J. Gen. Intern. Med., vol. 40, no. 1, pp. 127–134, 2025. doi: 10.1007/s11606-024-09050-9.

[48] K. F. Stanger-Hall, "Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes," CBE—Life Sci. Educ., vol. 11, no. 3, pp. 294–306, 2012. doi: 10.1187/cbe.11-11-0100.

[49] J. Sweller, "The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion," Educ. Psychol. Rev., vol. 35, Art. 87, 2023. doi: 10.1007/s10648-023-09817-2.

[50] J. Sweller, P. Ayres, and S. Kalyuga, Cognitive Load Theory. New York, NY, USA: Springer, 2011. doi: 10.1007/978-1-4419-8126-4.

[51] L. Y. Tan et al., "Artificial intelligence-enabled adaptive learning platforms: A review," Comput. Educ. Artif. Intell., vol. 9, Art. 100429, 2025. doi: 10.1016/j.caeai.2025.100429.

[52] M. Tarrant and J. Ware, "Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments," Med. Educ., vol. 42, no. 2, pp. 198–206, 2008. doi: 10.1111/j.1365-2923.2007.02957.x.

[53] G. Tsaparlis, "Higher and lower-order thinking skills: The case of chemistry revisited," J. Baltic Sci. Educ., vol. 19, no. 3, pp. 467–483, 2020. doi: 10.33225/jbse/20.19.467.

[54] B. Taylor, F. Kisby, and A. Reedy, "Rubrics in higher education: An exploration of undergraduate students' understanding and perspectives," Assess. Eval. Higher Educ., vol. 49, no. 6, pp. 978–992, 2024. doi: 10.1080/02602938.2023.2299330.

[55] W. J. van der Linden and C. A. W. Glas, Eds., Elements of Adaptive Testing. New York, NY, USA: Springer, 2010. doi: 10.1007/978-0-387-85461-8.

[56] W. J. van der Linden and R. K. Hambleton, Eds., Handbook of Modem Item Response Theory. New York, NY, USA: Springer, 1997.

[57] H. Vermeiren et al., "Balancing stability and flexibility: Investigating a dynamic K value approach for the Elo rating system in adaptive learning environments," User Model. User-Adapted Interact., vol. 36, Art. 4, 2026. doi: 10.1007/s11257-025-09439-z.

[58] H. Vermeiren et al., "Psychometrics of an Elo-based large-scale online learning system," Comput. Educ. Artif. Intell., vol. 8, Art. 100376, 2025. doi: 10.1016/j.caeai.2025.100376.

[59] E. Virranmäki, K. Valta-Hulkkonen, and A. Pellikka, "Geography tests in the Finnish Matriculation Examination in Paper and digital forms: An analysis of questions based on revised Bloom's taxonomy," Stud. Educ. Eval., vol. 66, Art. 100896, 2020. doi: 10.1016/j.stueduc.2020.100896.

[60] C. Voigt, L. Blömer, and U. Hoppe, "Analysing the cumulative hierarchy of the taxonomy of learning objectives in flipped classroom," Int. J. Higher Educ. Pedagogies, vol. 1, no. 1, pp. 1–12, 2020. doi: 10.33422/ijhep.v1i1.13.

[61] H. Xie et al., "Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017," Comput. Educ., vol. 140, Art. 103599, 2019. doi: 10.1016/j.compedu.2019.103599.

[62] A. C. M. Yang, B. Flanagan, and H. Ogata, "Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning," Comput. Educ. Artif. Intell., vol. 3, Art. 100104, 2022. doi: 10.1016/j.caeai.2022.100104.

[63] M. Yudelson et al., "Leveraging skill hierarchy for multi-level modeling with Elo rating system," in Proc. 6th ACM Conf. Learning @ Scale (L@S '19), 2019, pp. 1–4. doi: 10.1145/3330430.3333645.

[64] N. L. B. Zaidi et al., "Pushing critical thinking skills with multiple-choice questions: Does Bloom's taxonomy work?" Acad. Med., vol. 93, no. 6, pp. 856–859, 2018. doi: 10.1097/ACM.0000000000002087.

[65] Y. Zhang et al., "Development of a computerized adaptive testing for internet addiction," Frontiers Psychol., vol. 10, Art. 1010, 2019. doi: 10.3389/fpsyg.2019.01010.

APPENDIX A. SUMMARY OF NOTATION

TABLE III. SUMMARIZES THE PRINCIPAL NOTATION USED THROUGHOUT THE FORMAL SPECIFICATION OF THE HBC-GLICKO ARCHITECTURE. SYMBOLS ARE GROUPED BY FUNCTIONAL CATEGORY. SUBSCRIPT CONVENTIONS FOLLOW STANDARD PSYCHOMETRIC AND RATING-SYSTEM USAGE

Category	Symbol	Definition
General indices	n	Learner index
	t	Rating period (session) index
	b	Cognitive band index, $b \in \{1, 2, 3\}$

	$\mathcal{B}$	Set of cognitive bands {Low, Mid, High}
	$\mathcal{K}$	Set of Bloom categories within band b
Learner state	$r_{n,b,t}$	Rating of learner n in band b at period t
	$RD_{n,b,t}$	Rating deviation (uncertainty) of learner n in band b at period t
	$\boldsymbol{S}_{n,t}$	Full learner measurement state at period t
Item parameters	d_i	Calibrated difficulty of item i on the unified scale
	$RD_{i,t}$	Rating deviation of item i at period t
Response model	q	Scaling constant, q = ln(10) / 400
	$g(RD)$	Uncertainty weighting function
	$E_{n,i,t}$	Expected score of learner n on item i at period t
	$S_{n,i,t}$	Observed score, S ∈ [0, 1]
	$\delta_{n,i,t}$	Prediction error, S − E
	$\omega_{n,b,t}$	Glicko-style adaptive update weight
Within-band routing	$\pi_b(k)$	Target sampling distribution over categories in band b
	$c_t(k)$	Delivery counter for category k at step t
	$\Delta_t(k)$	Coverage deficit for category k at step t
	p^*	Target performance level for adaptive item selection
	H_s	Set of items already scored in session s
Progression	A_{b+1}	Entry anchor set for band b + 1
	d_{b+1}^{ref}	Mean reference difficulty of entry anchor items for band b + 1
	$p_{b\to b+1}$	Target transition probability from band b to band b + 1
	$T_{b\to b+1}$	Progression threshold from band b to band b + 1
Decision rules	z_b	Confidence multiplier for band b
	$LCB_{n,b}$	Lower confidence bound, r − z · RD
	$UCB_{n,b}$	Upper confidence bound, r + z · RD
	$N_{n,b}(k)$	Scored administrations for learner n in category k of band b
	m_b	Minimum evidence per category required for promotion
Inactivity	$t_{n,b}$	Sessions since last scored attempt by learner n in band b
	c_b	Inactivity-driven uncertainty growth rate for band b
	RD_{max}	Maximum admissible rating deviation
	$RD_{n,b}^{pre*}$	Effective pre-update uncertainty after inactivity adjustment

A Secure Chaotic DNA-Based Framework for Satellite Image Encryption

Asim Seedahmed Ali Osman[1]*, Ibrahim Rizqallah Alzahrani[2],
Aeshah Khalil I Alotaibi[3], Adil Mahmoud Mohamed Mahmoud[4], Daifallah Zaid Alotaibe[5]
Department of Software Engineering-College of Computer Science & Engineering,
University of Hafr Al Batin, Kingdom of Saudi Arabia[1]
Department of Computer Science and Engineering-College of Computer Science and Engineering,
University of Hafr Al Batin Kingdom of Saudi Arabia[2]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Kingdom of Saudi Arabia[3]
Faculty of Computer Science and Information Technology, White Nile University, Kosti, White Nile, Sudan[4]
Department of Software Engineering-College of Computer Science & Engineering,
University of Hafr Al Batin, Kingdom of Saudi Arabia[5]

***Abstract*—With the growing use of satellite imaging in environmental monitoring, defense, and remote sensing applications, protecting satellite images during storage and transmission has become increasingly important. Unlike textual data, satellite images contain high spatial redundancy and strong correlations among neighboring pixels, which can limit the effectiveness of conventional encryption methods when applied directly. To address this issue, this study presents a chaotic DNA-based framework for satellite image encryption. The proposed approach combines logistic-map-based chaotic key generation with DNA-inspired encoding and XOR operations to enhance pixel-level confusion and diffusion. The method was evaluated on a grayscale satellite image using common statistical security measures, including information entropy, adjacent-pixel correlation, NPCR, and UACI. The reported results indicate that the encrypted image achieved high entropy, low correlation between adjacent pixels, and strong sensitivity to small changes in the input image. These findings suggest that the proposed framework provides a reasonable basis for secure satellite image encryption. Although the current evaluation is limited in scope, the method shows encouraging performance and offers a useful direction for further investigation on larger and more diverse satellite image datasets.**

Keywords*—Chaotic cryptography; DNA computing; satellite image security; image encryption; information security; secure image transmission

I. Introduction

Satellite images have become an essential resource in many fields, including environmental monitoring, disaster response, military reconnaissance, and urban planning [15]. As satellite sensing technologies continue to advance, both the volume and resolution of captured images have increased significantly. This rapid growth has created a corresponding need for secure storage and transmission mechanisms to protect sensitive image data [17] from unauthorized *access* and misuse [1],[19]. Although conventional encryption algorithms such as AES and RSA provide strong security for general digital data [22], they are not always well suited to the structural characteristics of image data, particularly the high spatial redundancy and strong correlation among adjacent pixels. For this reason, their direct application to large image datasets may lead to practical limitations in efficiency and performance in some image-encryption settings [2, 12].

To address these limitations, chaos-based encryption methods have received considerable attention because of their sensitivity to initial conditions, nonlinear dynamic behavior, and ability to generate pseudo-random sequences. These characteristics make chaotic systems attractive for designing image-encryption schemes that require high unpredictability and strong key sensitivity [3],[24],[25]. In parallel, DNA-based cryptographic techniques have emerged as a promising approach for image protection by representing binary data using nucleotide symbols such as A, C, G, and T and performing logical operations in this encoded form. This representation can improve the confusion and diffusion properties of the encryption process, thereby enhancing the overall complexity of the protected image data [1],[4],[5].

Recent studies have explored the integration of chaotic systems with DNA-inspired operations for image encryption. Such approaches have shown encouraging results in terms of statistical security indicators, including improved entropy and reduced correlation among neighboring pixels in encrypted images [4],[5],[23]. However, the effectiveness of these methods still depends on several important factors, including the clarity of the encryption procedure, the reliability of the key-generation mechanism, and the scope of the experimental validation used to support the reported security claims.

Motivated by these considerations, this study presents a chaotic DNA-based framework for satellite image encryption. The proposed approach combines logistic-map-based chaotic key generation with DNA-inspired encoding and XOR operations to strengthen pixel-level transformation during the encryption process. In the current study, the framework is examined as a proof-of-concept for grayscale satellite image encryption. The main contributions of this work can be summarized as follows:

- A chaotic DNA-based framework for satellite image encryption.

- A logistic-map-driven key-generation process for generating encryption sequences.
- The use of DNA-inspired encoding and XOR-based operations to enhance confusion and diffusion at the pixel level.
- A preliminary statistical evaluation based on entropy, correlation, NPCR, and UACI.

The remainder of this study is organized as follows. Section II reviews the related literature. Section III presents the proposed methodology. Section IV describes the experimental setup and reports the results. Section V discusses the findings and limitations of the study. Finally, Section VI concludes the study and outlines directions for future work.

II. Related Work

A. Image Encryption: Challenges Beyond Traditional Methods

Securing the transmission of digital images is increasingly vital, particularly in satellite imaging, where both the sensitivity and sheer size of data are significant concerns. While established encryption methods such as AES and RSA offer strong protection for conventional data, they may not always be well suited for image encryption. Their intensive computational demands and inability to handle image-specific properties—like high redundancy and strong pixel correlation—make them less effective in this context [2],[12]. Furthermore, these techniques often fail to manage the large file sizes typical of satellite imagery, especially when real-time processing is required.

B. Chaos Theory in Image Encryption

Chaos-based encryption methods have received considerable attention in the field of image security because of their sensitivity to initial conditions, nonlinear dynamic behavior, and ability to generate pseudo-random sequences. These properties make chaotic systems well-suited for image-encryption tasks, where unpredictability and key sensitivity play an important role. Various chaotic models, including logistic maps, tent maps, the Henon system, and piecewise linear chaotic maps (PWLCM), have been used in previous studies to generate key streams and support permutation and diffusion processes in encrypted images [3], [6],[24].

In addition, chaos-based approaches are often considered attractive because they can increase the complexity of the encryption process while maintaining relatively low computational cost [18]. This makes them suitable for applications that require efficient image protection, particularly in settings where processing speed and limited computational resources are important [6]. Previous studies have reported that combining chaotic mechanisms can further improve the complexity of image-encryption schemes and enhance their resistance to unauthorized recovery attempts [24, 25]. Overall, chaos-based encryption remains a widely studied approach in image security due to its flexibility, efficiency, and strong dependence on initial parameter values.

C. DNA Cryptography in Multimedia Protection

DNA-based cryptographic techniques have attracted growing attention in multimedia security because they provide an alternative way to represent and process digital data. In this approach, binary information is encoded using nucleotide symbols such as A, T, C, and G, and a set of logical operations can then be performed within this encoded domain. Early DNA-based encryption models often relied on fixed encoding rules, whereas more recent studies have introduced dynamic DNA coding strategies and logical operations such as XOR, addition, and subtraction, often controlled by chaotic sequences [4], [5], [14].

Previous studies have shown that combining DNA-based data representation with chaotic control can improve image-encryption performance and strengthen statistical security properties [5], [7]. Previous studies reported that DNA-based operations can contribute to both confusion and diffusion during the encryption process [5], [23], making the encrypted image less predictable and more resistant to statistical analysis. These methods have also been associated with reduced correlation among adjacent pixels and improved entropy values in encrypted images, which are widely used indicators in image-security evaluation.

D. Hybrid Chaos-DNA Image Encryption

The combination of chaos and DNA encryption has been proposed as a way to address the limitations of each approach individually. Previous studies introduced a chaos-driven dynamic DNA coding method that adapts encoding rules based on chaotic sequences, ensuring per-pixel variability and greater unpredictability [8], [23].

Hybrid systems apply chaotic maps for key stream generation and pixel permutation, while DNA encoding and logical operations provide data transformation at the nucleotide level [3], [9]. This dual-layer approach disrupts statistical patterns within the image, achieving superior encryption quality compared to single-domain techniques [10, 24].

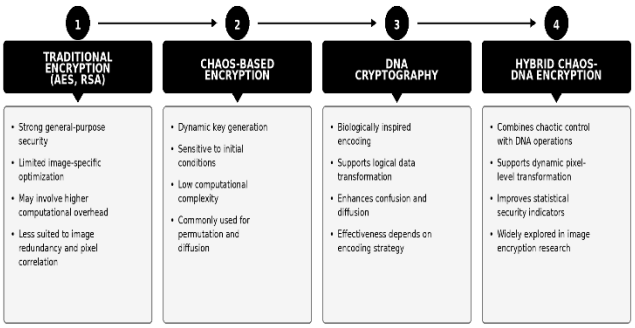


Fig. 1. Overview of conventional, chaos-based, DNA-based, and hybrid image encryption approaches.

E. Gaps in Existing Literature

Although chaos-DNA image-encryption methods have shown promising results in previous studies, several limitations remain in the existing literature. First, many reported works focus on general image datasets rather than satellite imagery as a specific application context [2, 8]. Second, several studies emphasize statistical security indicators without providing sufficiently clear methodological descriptions or broad experimental validation. In addition, the reported performance of many existing methods depends strongly on the design of the

key-generation process and the way DNA-based operations are integrated into the encryption framework [4], [13]. These observations indicate the need for further studies that provide clearer encryption procedures and a more focused evaluation for satellite image encryption.

F. Motivation for the Proposed Framework

Motivated by the above observations (Fig. 1), this study proposes a chaotic DNA-based framework for satellite image encryption. The framework is intended to combine logistic-map-based chaotic key generation with DNA-inspired encoding and XOR operations in order to strengthen pixel-level transformation during encryption. The main motivation behind this design is to explore a simple and structured encryption approach that can benefit from the sensitivity of chaotic sequences and the representational flexibility of DNA-based operations.

In this study, the proposed framework is examined as a proof-of-concept for grayscale satellite image encryption. The work is intended to provide a clearer methodological basis for integrating chaotic key generation with DNA-based processing, while offering a preliminary statistical evaluation using common image-security indicators.

III. Proposed Methodology

This section presents the design and operational workflow of the proposed encryption framework for satellite image protection. The framework is intended to combine chaotic key generation with DNA-inspired data transformation in a structured and lightweight encryption process. The main objective is to enhance pixel-level confusion and diffusion while maintaining a clear methodological design that can be applied to grayscale satellite image encryption. In this study, the framework is examined as a proof-of-concept setting for secure image encryption using common statistical security indicators.

A. Overall Architecture

The proposed encryption framework is organized around three main components:

1) Preprocessing module: This stage prepares the input satellite image in a suitable format for the encryption process. In the current implementation, the framework is applied to grayscale image data to ensure a consistent input structure for subsequent processing steps.

2) Chaotic key-generation module: This component uses the logistic map to generate pseudo-random sequences that serve as the basis for the encryption process. Because chaotic systems are highly sensitive to initial conditions [3], [24], the generated sequences introduce unpredictability and strengthen the key-dependent behavior of the framework [3], [25].

3) DNA-based encryption module: In this stage, image data are transformed using DNA-inspired encoding and XOR operations. These operations are applied to increase confusion and diffusion at the pixel level [4], [5], [10] and to support the overall complexity of the encryption process.

Fig. 2 illustrates the overall workflow of the proposed framework for satellite image encryption. The framework consists of sequential stages that begin with image preprocessing, followed by logistic-map-based key generation, DNA encoding, and XOR-based transformation, and then proceed to the corresponding decryption steps for image recovery.

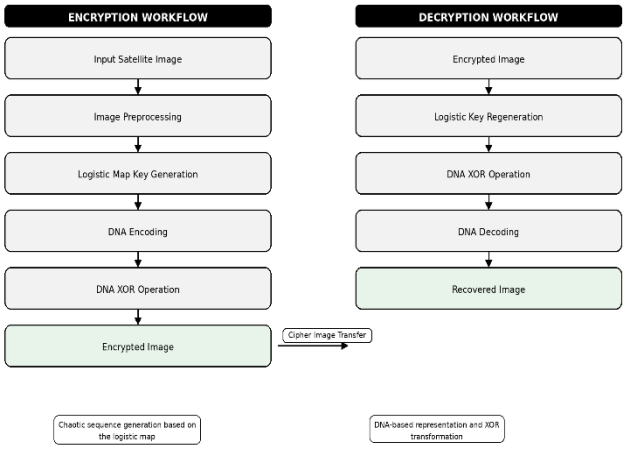


Fig. 2. Overall workflow of the proposed chaotic DNA-based satellite image encryption framework.

As shown in Fig. 2, the encryption process starts with preparing the input satellite image in a suitable format for further processing. A pseudo-random key sequence is then generated using the logistic map and used to support the encryption procedure. After that, the image data are transformed through DNA-based encoding and XOR operations to produce the encrypted image. For decryption, the same key-generation mechanism is used to regenerate the required sequence, after which inverse XOR and DNA decoding steps are applied to recover the original image. This workflow provides a clear representation of the main operational stages of the proposed framework.

B. Chaotic Key Generation

Chaotic key generation is a central component of the proposed framework because it introduces sensitivity to the initial parameters and supports the production of pseudo-random sequences for encryption. In this study, the logistic map is used as the chaotic generator because of its simple formulation and well-known nonlinear behavior. The logistic map is defined as Eq. (1).

$$x_{(n+1)} = \mu x_n(1 - x_n) \quad (1)$$

where, x_n in (0,1) denotes the chaotic state at iteration n, and \mu is the control parameter. When \mu is chosen in the chaotic region, the generated sequence becomes highly sensitive to small changes in the initial value x_0, which makes it suitable for encryption-oriented key generation.

For an input image of size M x N, the logistic map is iterated to generate a sequence of length M x N. The resulting real-valued chaotic sequence is then transformed into an integer-valued key stream that can be used in the subsequent encryption stages. A practical form of this transformation is given in Eq. (2).

$$k_i = \lfloor 256x_i \rfloor,\ i = 1,2,\ldots,MN \quad (2)$$

where, k_i represents the generated key value corresponding to the chaotic state x_i. This step maps the continuous chaotic output to discrete values that are compatible with digital image processing.

The generated key stream is then used to support the DNA-based encryption stage. Because the logistic map is highly sensitive to the initial condition x_0 and the control parameter \mu, even a small change in these values leads to a different chaotic sequence and, consequently, a different encryption result. This sensitivity strengthens the key-dependent behavior of the proposed framework and contributes to the unpredictability of the encryption process.

C. DNA Encoding and Logical Operations

After generating the chaotic key sequence, the framework applies DNA-inspired encoding to the image data. In DNA-based image encryption, binary values are represented using nucleotide symbols, namely A, T, C, and G. This representation provides an additional transformation layer and allows logical operations to be performed in the encoded domain.

In the proposed framework, each image pixel is first converted into its 8-bit binary form and then mapped into a DNA sequence according to a predefined encoding rule. This process can be expressed as Eq. (3).

$$D_i = E(p_i) \tag{3}$$

where, (p_i) denotes the original pixel value, (E(\cdot)) represents the DNA encoding function, and (D_i) is the DNA-encoded representation of that pixel.

After DNA encoding, the generated chaotic key sequence is transformed into a compatible form and combined with the encoded image data through a DNA-based XOR operation. This step increases the complexity of the encryption process and enhances pixel-level confusion. The operation can be written as Eq. (4).

$$C_i = D_i \oplus K_i \tag{4}$$

where, (K_i) is the encoded key element generated from the chaotic sequence, and (C_i) is the resulting DNA-coded encrypted value after the XOR operation.

Once this step is completed, the resulting DNA sequence is converted back into binary form and then reconstructed into pixel values to produce the encrypted image. During decryption, the same chaotic key sequence is regenerated, and the inverse procedure is applied to recover the original image. In this way, the DNA-based stage contributes to both confusion and diffusion within the proposed encryption framework.

D. Encryption Process Flow

The encryption process of the proposed framework is carried out through a sequence of structured stages that combine chaotic key generation with DNA-based data transformation. The purpose of this process is to convert the original satellite image into an encrypted form with reduced statistical predictability and stronger pixel-level confusion.

First, the input satellite image is prepared in a suitable format for encryption. In the current implementation, the image is processed as a grayscale image to ensure consistency in the subsequent stages. After preprocessing, the logistic map is iterated using the selected initial condition and control parameter to generate a chaotic sequence. This sequence is then transformed into a key stream that can be used in the encryption stage.

Next, the image pixel values are converted into binary form and encoded according to the selected DNA representation rule. The generated chaotic key stream is also transformed into a compatible encoded form. A DNA-based XOR operation is then applied between the encoded image data and the corresponding key sequence. This step introduces additional transformation at the pixel level and contributes to the complexity of the encrypted output.

After the XOR operation, the resulting DNA sequence is converted back into binary form and reconstructed into pixel values. These values form the encrypted image. Through this sequence of operations, the framework combines the sensitivity of chaotic key generation with the symbolic transformation provided by DNA-based encoding.

For clarity, the encryption procedure can be summarized as follows:

- Read the input satellite image.
- Preprocess the image and represent it in grayscale form.
- Generate a chaotic sequence using the logistic map.
- Convert the chaotic sequence into a key stream suitable for encryption.
- Encode the image pixels using the selected DNA rule.
- Apply the DNA-based XOR operation between the encoded image data and the encoded key sequence.
- Decode the resulting sequence and reconstruct the encrypted image.

E. Decryption Process Flow

The decryption process is carried out by reversing the main stages of the encryption procedure. To recover the original image correctly, the same initial condition and control parameter used in the logistic map during encryption must be applied again in the decryption stage. This allows the chaotic key sequence to be regenerated in an identical form.

First, the received encrypted image is converted into a suitable format for decryption. The logistic map is then re-initialized using the same secret parameters in order to regenerate the corresponding chaotic key sequence. After that, the encrypted image data is represented in DNA form using the same encoding rule adopted during the encryption process.

Next, the regenerated key sequence is transformed into a compatible encoded form and used in the DNA-based XOR operation. Because the same key sequence is reproduced, this step reverses the transformation applied during encryption. The resulting DNA sequence is then decoded back into binary form and reconstructed into pixel values to recover the original image.

For clarity, the decryption procedure can be summarized as follows:

- Read the received encrypted image.
- Regenerate the chaotic key sequence using the logistic map and the original secret parameters.
- Encode the encrypted image data using the same DNA rule used in encryption.
- Apply the DNA-based XOR operation using the regenerated key sequence.
- Decode the resulting DNA sequence and reconstruct the original pixel values.
- Recover the original satellite image.

Accurate synchronization of the logistic-map parameters and the DNA encoding rule between the sender and the receiver is essential for successful decryption and correct recovery of the original image data.

F. Key Features of the Framework

The proposed framework includes several characteristics that distinguish it from conventional image-encryption approaches. First, it combines chaotic key generation with DNA-inspired data transformation within a single encryption workflow. This integration provides a structured mechanism for introducing key sensitivity and pixel-level transformation during encryption.

Second, the framework is designed in a relatively simple and interpretable manner. The use of the logistic map enables pseudo-random sequence generation through a compact mathematical model, while DNA-based encoding and XOR operations provide an additional transformation layer for image data. This combination supports the overall complexity of the encryption process without relying on an overly complicated design.

Third, the framework is developed as a proof-of-concept for grayscale satellite image encryption. In this context, it is intended to demonstrate how chaotic sequence generation and DNA-based representation can be integrated within a clear methodological structure and evaluated using common statistical image-security indicators.

Overall, the proposed framework provides a lightweight and structured approach for exploring satellite image encryption using chaotic and DNA-based operations [2, 8].

IV. Experimental Setup and Results

This section describes the experimental setting used to evaluate the proposed chaotic DNA-based encryption framework. It presents the image preparation procedure, the implementation setting, and the statistical measures used to assess the encryption results. The evaluation is intended as a proof-of-concept analysis of the proposed method on grayscale satellite image data. The obtained results are reported and discussed in terms of commonly used image-security indicators, including entropy, adjacent-pixel correlation, NPCR, and UACI.

A. Experimental Environment

The experimental evaluation was conducted on a Windows 11 Pro (64-bit) platform equipped with an Intel Core i7-11800H processor and 16 GB of RAM. The proposed encryption framework was implemented in Python 3.11 using the Anaconda environment.

The main Python libraries used in the implementation were as follows:

- NumPy: used for numerical computations and the generation of pseudo-random sequences required in the chaotic processing stage.
- OpenCV: used for image reading, writing, and pixel-level image manipulation.
- Matplotlib: used for visualizing histograms and correlation plots.
- SciPy: used for entropy calculation and related statistical analysis.

B. Dataset Preparation and Preprocessing

A publicly available grayscale satellite image of size 256 × 256 pixels was used in the experimental evaluation. This image was selected to provide a proof-of-concept setting for examining the behavior of the proposed encryption framework on satellite image data. Fig. 3 shows the original input image used in the experiments.

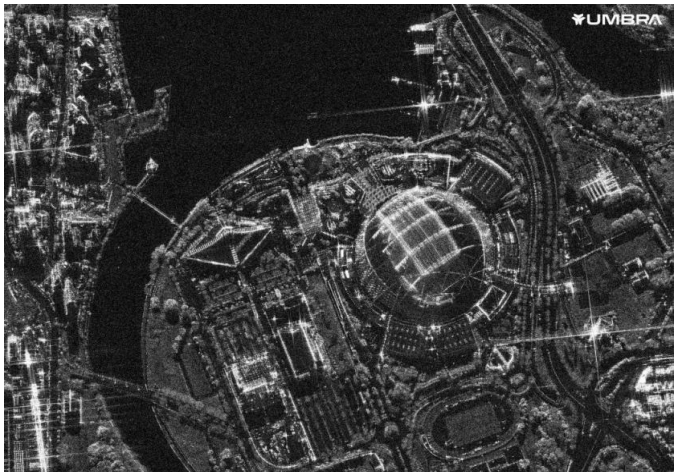


Fig. 3. Original grayscale satellite image used for encryption experiments.

Before encryption, the image was preprocessed using OpenCV to ensure a consistent input format for the subsequent stages of the proposed framework.

The preprocessing stage involved reading the input image in grayscale mode and resizing it to 256 × 256 pixels. This step ensured consistency in the input representation before applying the encryption procedure.

C. Chaotic Key Generation

A logistic map was used to generate the chaotic sequence employed in the key-generation stage. The mathematical formulation of the logistic map and its transformation into a discrete key stream were presented earlier in Section III-B. In the experimental implementation, the generated chaotic sequence was used to construct the key stream required for the encryption process.

D. DNA Encoding and Logical Operations

In this stage, binary pixel values were mapped to DNA sequences, and DNA-based XOR operations were applied to support pixel-level transformation during encryption. The mathematical formulation of the DNA encoding and XOR operations was introduced earlier in Section III-C. In the implementation, each pixel value was converted into its 8-bit binary form, divided into four pairs of bits, and mapped using the adopted DNA rule: 00 → A, 01 → C, 10 → G, and 11 → T. After that, the encoded image data were combined with the encoded chaotic key sequence through a DNA-based XOR operation. The resulting DNA sequence was then converted back into binary form and reconstructed into pixel values to produce the encrypted image.

E. Encryption Process

In the implementation, the encryption procedure was applied to each pixel of the input image using the generated chaotic key stream. For every pixel, the binary value of the image data was converted into DNA form, and the corresponding key element was transformed into a compatible DNA representation. A DNA-based XOR operation was then applied between the encoded pixel value and the encoded key value. The resulting DNA sequence was subsequently converted back into binary form and reconstructed into the encrypted pixel value. This process was repeated over the entire image to produce the final encrypted image.

Fig. 4 shows the encrypted output generated by the proposed framework after applying chaotic key generation, DNA encoding, and DNA-based XOR transformation to the input satellite image.

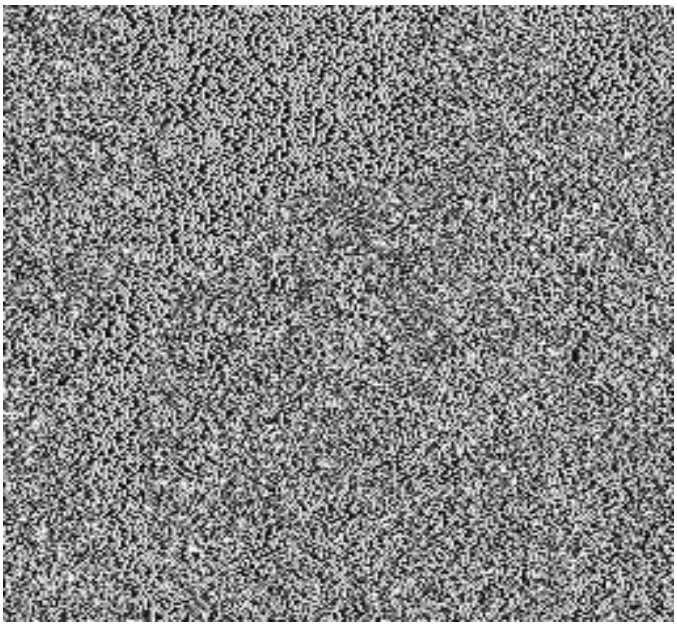


Fig. 4. Encrypted image produced by the proposed framework.

As shown in Fig. 4, the encrypted image exhibits a noise-like appearance with no visually recognizable structures from the original image. This visual behavior indicates that the proposed framework was able to significantly alter the spatial distribution of pixel intensities in the encrypted output.

F. Evaluation Metrics

The encryption performance was evaluated using standard statistical security metrics commonly employed in image-encryption studies.

The evaluation focused on four widely used indicators: information entropy, adjacent-pixel correlation, NPCR, and UACI. Information entropy was used to assess the randomness of the encrypted image distribution. Adjacent-pixel correlation was used to measure the relationship between neighboring pixels in the horizontal, vertical, and diagonal directions. NPCR was used to quantify the percentage of differing pixels between two images, while UACI was used to measure the average intensity difference between them. Together, these metrics provide a statistical view of the encryption behavior of the proposed framework.

Although these metrics are widely used in image-encryption analysis, they do not by themselves provide a complete proof of security against all possible attack models. In this study, they are used as standard indicators for a preliminary statistical evaluation of the proposed framework.

G. Results and Analysis

The experimental results indicate that the proposed framework produced statistically encouraging encryption behavior for the tested satellite image. The entropy of the encrypted image reached 7.99, which is close to the ideal value of 8 and suggests a high level of randomness in the encrypted output [1],[5],[23]. In addition, the NPCR value reached 99.61%, indicating a high percentage of pixel changes between the compared images. The UACI value was 33.32%, which is consistent with the expected behavior of a strongly transformed encrypted image. Furthermore, the correlation values in the horizontal, vertical, and diagonal directions were close to zero, indicating a substantial reduction in the linear dependency among adjacent pixels after encryption [5],[7],[23].

Overall, these results suggest that the proposed framework was effective in altering the statistical structure of the original image and producing an encrypted output with noise-like characteristics. Table I summarizes the main performance metrics obtained in the experimental evaluation.

TABLE I. PERFORMANCE METRICS OF THE PROPOSED ENCRYPTION FRAMEWORK

Metric	Result
Entropy (Encrypted)	7.99
NPCR (%)	99.61
UACI (%)	33.32
Correlation (H, V, D)	-0.003, 0.005, -0.002

V. SIMULATION RESULTS AND SECURITY EVALUATION

This section presents a statistical evaluation of the proposed chaotic DNA-based encryption framework, with emphasis on the encryption performance of the tested satellite image. The analysis was carried out using standard image-security indicators and implementation-based observations. A publicly available grayscale satellite image of size 256 × 256 pixels was used in the experiments. The implementation was performed on

a Windows 11 system equipped with an 11th Gen Intel Core i7 processor and 16 GB of RAM, using Python 3.11 within the Anaconda environment.

A. Ability to Resist Statistical Attacks

1) Histogram analysis: A common objective of image encryption is to reduce visible redundancy in pixel-intensity distributions and make the encrypted image less predictable than the original image. As shown in Fig. 5, the histogram of the original satellite image exhibits noticeable variations in intensity distribution, reflecting the structural patterns present in the plain image. By contrast, the histogram of the encrypted image appears more dispersed across the grayscale range, with no obvious dominant peaks comparable to those observed in the original image. This visual change suggests that the proposed framework altered the statistical distribution of pixel intensities in the encrypted output. However, histogram analysis alone is not sufficient to establish complete statistical security, and it should be interpreted together with quantitative indicators such as entropy and adjacent-pixel correlation.

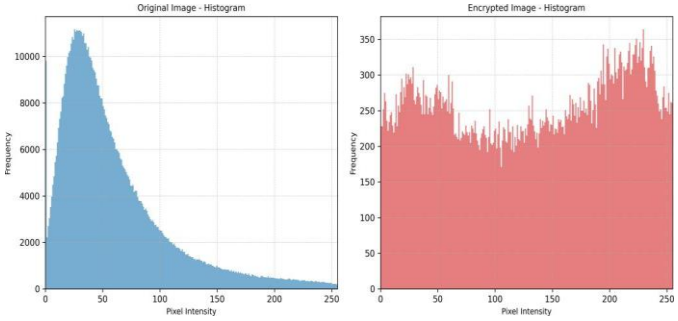


Fig. 5. Histogram comparison between the original and encrypted satellite images: (a) Histogram of the original image, (b) Histogram of the encrypted image.

2) Correlation analysis: To evaluate the statistical relationship between neighboring pixels, correlation coefficients were calculated in the horizontal, vertical, and diagonal directions. In the original satellite image, relatively high correlation values were observed, reflecting the spatial redundancy that is typically present in plain image data. After encryption, these values were reduced substantially and became close to zero, indicating a strong reduction in linear dependency among adjacent pixels.

Fig. 6-8 presents scatter plots for the original and encrypted images and illustrates the visual change in pixel correlation before and after encryption. In the original image, the points appear more strongly clustered, reflecting the spatial dependency among neighboring pixels. In contrast, the encrypted image shows a substantially more dispersed distribution, indicating a marked reduction in linear dependency after encryption. Although some local clustering can still be observed in the encrypted scatter plots, they remain considerably less structured than those of the original image. This behavior is consistent with the numerical correlation results reported in Table II and suggests that the proposed framework significantly reduced pixel-level statistical dependency in the encrypted output.

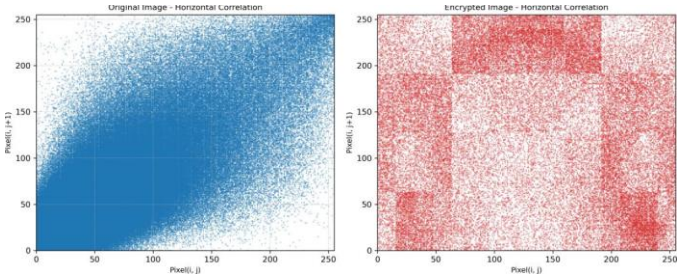


Fig. 6. Scatter plot comparison of horizontally adjacent pixels in the original and encrypted satellite images, showing the reduction in pixel correlation after encryption.

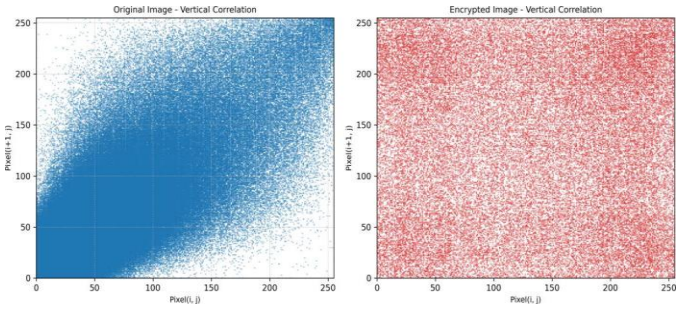


Fig. 7. Scatter plot comparison of vertically adjacent pixels in the original and encrypted satellite images, showing the reduction in pixel correlation after encryption.

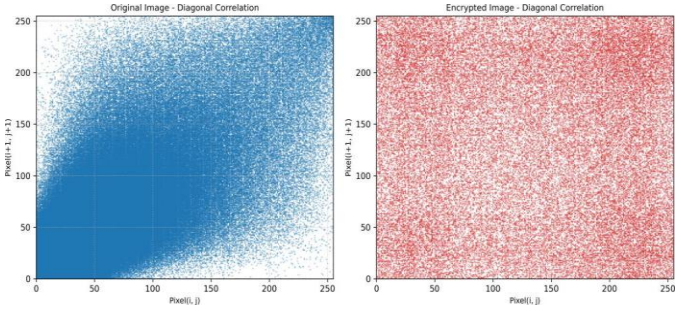


Fig. 8. Scatter plot comparison of diagonally adjacent pixels in the original and encrypted satellite images.

TABLE II. CORRELATION COEFFICIENTS OF ADJACENT PIXELS BEFORE AND AFTER ENCRYPTION

Image	Plain Image			Encrypted Image		
	Horizontal	*Vertical*	*Diagonal*	*Horizontal*	*Vertical*	*Diagonal*
Satellite Image	0.96	0.95	0.94	-0.003	0.005	-0.002

Table II reports the correlation coefficients of adjacent pixels in the horizontal, vertical, and diagonal directions before and after encryption. The original image shows high correlation values, which reflect the strong spatial dependency typically present in plain image data. After encryption, the corresponding values become very close to zero, indicating a substantial reduction in linear dependency among neighboring pixels.

3) Information entropy: Information entropy is commonly used to measure the degree of randomness or unpredictability in an encrypted image. For an 8-bit grayscale image, the ideal entropy value is 8. In this study, the encrypted image achieved

an entropy value of 7.99, which is very close to the ideal value. By comparison, the entropy of the plain satellite image was 7.55. This result suggests that the proposed framework increased the randomness of the image data after encryption and supports the statistical effectiveness of the encryption process. The corresponding entropy values are reported in Table III.

TABLE III. INFORMATION ENTROPY OF THE PLAIN AND ENCRYPTED SATELLITE IMAGES

Image	Plain Image Entropy	Encrypted Image Entropy
Satellite Image	7.55	7.99

B. Resistance to Differential Attacks

Differential attack analysis is commonly used to evaluate the sensitivity of an encryption method to small changes in the input image. In this study, the Number of Pixel Change Rate (NPCR) and the Unified Average Changing Intensity (UACI) were used to assess the effect of pixel-level changes on the encrypted output. The experimental results showed an NPCR value of 99.61% and a UACI value of 33.35%, indicating that the proposed framework produced substantial changes in the encrypted image in response to small variations in the input. These results suggest a high degree of sensitivity to differential changes [1] and support the statistical effectiveness of the encryption process. A summary of the obtained results is presented in Table IV.

TABLE IV. NPCR AND UACI RESULTS FOR DIFFERENTIAL ATTACK ANALYSIS

Image	NPCR (%)	UACI (%)
Satellite Image	99.61	33.35

C. Key Sensitivity Analysis

An important property of the proposed framework is its sensitivity to the secret parameters used in the chaotic key-generation stage. Because the logistic map is highly dependent on its initial condition and control parameter, even a very small change in these values leads to a different chaotic sequence and, consequently, a different encrypted output.

In the present framework, the parameters (x_0) and (\mu) must be reproduced accurately during decryption to regenerate the same key sequence used in the encryption stage. Any mismatch in these parameters prevents the correct recovery of the original image. This behavior indicates that the proposed method is strongly dependent on parameter synchronization and exhibits a high degree of key sensitivity [25].

From a security perspective, this sensitivity is important because it reduces the likelihood of recovering the correct image without the exact parameter settings. Therefore, the strength of the proposed framework is supported by the strong dependence of the encryption and decryption processes on the correct chaotic parameters.

D. Computational Complexity Analysis

The computational complexity of the proposed encryption framework can be expressed as $O(M \times N)$, where M and N denote the height and width of the input image, respectively. This complexity arises because the main encryption operations are applied at the pixel level across the entire image. In the current implementation, the framework combines logistic-map-based key generation with DNA encoding and XOR operations in a relatively simple processing structure.

From an implementation perspective, the framework does not rely on deeply nested transformations or computationally expensive optimization procedures. This makes the overall processing workflow relatively lightweight for the tested image size. However, the present evaluation is limited to a proof-of-concept setting on grayscale satellite imagery, and broader analysis on larger datasets and different computational settings remains an important direction for future work.

E. Summary of Security and Performance Metrics

A summary of the main security and performance metrics obtained for the proposed framework is presented in Table V.

TABLE V. SUMMARY OF THE PROPOSED FRAMEWORK'S SECURITY AND PERFORMANCE METRICS

Metric	Proposed Framework	Reference / Expected Behavior
Entropy (bits)	7.99	Close to 8.0
NPCR (%)	99.61	Greater than 99.5
UACI (%)	33.35	Around 33%
Correlation (H, V, D)	-0.003, 0.005, -0.002	Close to 0

VI. DISCUSSION

The experimental results indicate that the proposed chaotic DNA-based framework was able to alter the statistical characteristics of the tested satellite image and produce an encrypted output with increased randomness and reduced pixel-level dependency. The obtained values of entropy, correlation, NPCR, and UACI suggest that the framework achieved statistically encouraging encryption behavior under the experimental setting considered in this study. At the same time, these results should be interpreted within the scope of a proof-of-concept evaluation based on grayscale satellite image data and standard statistical image-security indicators.

A. Theoretical Contributions

From a methodological perspective, the proposed framework contributes to the integration of chaotic key generation with DNA-inspired image transformation within a relatively simple encryption structure. The logistic map provides a compact nonlinear mechanism for generating key-dependent pseudo-random sequences, while DNA-based encoding and XOR operations introduce an additional symbolic transformation layer during encryption. The combination of these two components supports pixel-level confusion and contributes to the overall complexity of the encrypted image representation.

The reported entropy, correlation, NPCR, and UACI results suggest that the proposed framework was effective in changing the statistical structure of the tested image after encryption. In this sense, the theoretical contribution of the study lies not in introducing an entirely new cryptographic primitive, but in

presenting a structured proof-of-concept framework that combines chaotic and DNA-based operations in a clear and interpretable way for satellite image encryption.

B. Practical Implications for Satellite Image Security

From an application perspective, the proposed framework illustrates how chaotic key generation and DNA-based transformation can be combined to protect satellite image data within a lightweight implementation structure. The present results suggest that the method can alter the visual and statistical characteristics of the input image in a way that makes the encrypted output less predictable than the original image.

However, the current implementation should be viewed as an initial proof-of-concept rather than a complete deployment-ready solution for operational satellite systems. Broader practical use would require additional validation on larger and more diverse image datasets, as well as further analysis under different computational and communication settings.

C. Comparison with Related Work

Relative to prior chaos-based and DNA-based image-encryption studies, the proposed framework follows a similar general direction by combining nonlinear key generation with symbolic image transformation. The statistical results obtained in this study, particularly in terms of entropy, adjacent-pixel correlation, NPCR, and UACI, are consistent with the types of indicators commonly reported in the related literature.

At the same time, the present work should be interpreted within its actual experimental scope. The proposed framework was evaluated as a proof-of-concept on grayscale satellite image data, and therefore the comparison with related work is best understood as a contextual rather than an exhaustive benchmark comparison. A broader and more rigorous comparative analysis against additional recent methods remains an important direction for future research.

D. Limitations and Future Work

Several limitations of the current study should be acknowledged. First, the evaluation was conducted on grayscale satellite imagery within a proof-of-concept setting, which limits the generalizability of the reported results. Second, the current analysis relied mainly on standard statistical security indicators such as entropy, correlation, NPCR, and UACI. Although these metrics are widely used in image-encryption studies [11], they do not by themselves provide a complete assessment against all possible attack models.

In addition, the present implementation did not include a broader evaluation on multispectral or high-dynamic-range satellite imagery, nor did it include extensive benchmarking against a large set of recent encryption baselines. Future work may therefore extend the framework to more diverse image types, incorporate additional robustness and comparative experiments, and explore alternative chaotic designs or more advanced DNA-based transformation strategies.

E. Broader Implications in Cryptography

The study also highlights the potential value of combining ideas from nonlinear dynamics and biologically inspired data representation in multimedia security research. Even within the limited scope of the present implementation, the integration of chaotic processing and DNA-based transformation demonstrates how cross-disciplinary concepts can be used to develop structured image-encryption frameworks [4],[8],[23].

More broadly, this line of research may support future work on lightweight and application-oriented cryptographic methods for image and multimedia protection [20], [21]. In this respect, the proposed framework serves as a useful exploratory step toward more comprehensive studies on secure image encryption using chaos-based and DNA-inspired techniques.

VII. Conclusion

This study presented a chaotic DNA-based framework for satellite image encryption that combines logistic-map-based key generation with DNA-inspired encoding and XOR operations. The proposed framework was developed as a proof-of-concept approach for grayscale satellite image encryption and was evaluated using common statistical security indicators.

The experimental results showed that the encrypted image achieved high entropy, low adjacent-pixel correlation, and encouraging NPCR and UACI values under the adopted evaluation setting. These findings suggest that the proposed framework was effective in altering the statistical characteristics of the tested image and producing an encrypted output with reduced predictability.

At the same time, the current study is limited to a proof-of-concept evaluation on grayscale satellite imagery and standard statistical image-security metrics. Therefore, the reported results should be interpreted within this scope. Future work may extend the framework to more diverse satellite image types, incorporate broader comparative and robustness analyses, and investigate alternative chaotic designs or more advanced DNA-based transformation strategies.

References

[1] Belazi, A., Talha, M., Kharbech, S., & Xiang, W. (2019). Novel Medical Image Encryption Scheme Based on Chaos and DNA Encoding. IEEE Access, 7, 36667–36681.

[2] Zhang, B., & Liu, L. (2023). Chaos-Based Image Encryption: Review, Application, and Challenges. Mathematics, 11(11), 2585.

[3] Chai, X., Fu, X., Gan, Z., Lu, Y., & Chen, Y. (2019). A color image cryptosystem based on dynamic DNA encryption and chaos. Signal Processing, 155, 44–62.

[4] Liu, L., Wang, D., & Lei, Y. (2020). An Image Encryption Scheme Based on Hyper Chaotic System and DNA With Fixed Secret Keys. IEEE Access, 8, 46400–46416.

[5] Zhang, Y. (2018). The Image Encryption Algorithm Based on Chaos and DNA Computing. Multimedia Tools and Applications, 77(16), 21589–21615.

[6] Yan, M., Liu, M., & Li, C. (2025). DNA Color Image Encryption Based on Conservative Chaotic System. Scientific Reports, 15(1), 9278.

[7] Liu, Y., Tang, J., & Xie, T. (2014). Cryptanalyzing a RGB image encryption algorithm based on DNA encoding and chaos map. Optics & Laser Technology, 60, 111–115.

[8] Patidar, V., & Kaur, G. (2023). A Novel Conservative Chaos Driven Dynamic DNA Coding for Image Encryption. Frontiers in Applied Mathematics and Statistics, 8, 1100839.

[9] Enayatifar, R., Guimaraes, F. G., & Siarry, P. (2019). Index-based permutation-diffusion in multiple-image encryption using DNA sequence. Optics and Lasers in Engineering, 115, 131–140.

[10] Ur Rehman, A., Liao, X., Ashraf, R., Ullah, S., & Wang, H. (2018). A color image encryption technique using exclusive-OR with DNA complementary rules based on chaos theory and SHA-2. Optik, 159, 348–367.

[11] Sokouti, M., & Sokouti, B. (2018). A systematic review on color image encryption using DNA properties. Computer Science Review, 29, 14–20.

[12] Panda, M. (2016). Performance analysis of encryption algorithms for security. Proceedings of SCOPES 2016, IEEE, 278–284.

[13] Li, C., Li, S., Asim, M., Nunez, J., Alvarez, G., & Chen, G. (2016). On the Security Defects of an Image Encryption Scheme. arXiv.

[14] Chattopadhyay, C., Sarkar, B., & Mukherjee, D. (2015). Encoding by DNA Relations and Randomization Through Chaotic Sequences for Image Encryption. arXiv.

[15] Galdran, A. (2018). Image Dehazing by Artificial Multiple-Exposure Image Fusion. Signal Processing, 149, 135–147.

[16] Vanishreeprasad, S., & Pushpalatha, K. N. (2015). Design and implementation of hybrid cryptosystem using AES and Hash Function. IOSR Journal of Electronics and Communication Engineering, 10(3), 18–24.

[17] Meenakumari, M., & Athisha, G. (2014). Improving message authentication by integrating encryption with hash function and its VLSI implementation.

[18] Thabit, F., Alhomdy, S., & Jagtap, S. (2021). Security analysis and performance evaluation of a new lightweight cryptographic algorithm for cloud computing.

[19] Gastermann, B., Stopper, M., Kossik, A., & Katalinic, B. (2015). Secure implementation of an on-premises cloud storage service for SMEs.

[20] Khan, S. S., & Tuteja, R. R. (2015). Security in cloud computing using cryptographic algorithms.

[21] Gong, Z., Nikova, S., Law, Y. W., Juels, A., & Paar, C. (2012). RFID Security and Privacy.

[22] Princy, P. (2015). A comparison of symmetric key algorithms DES, AES, Blowfish, RC4, RC6.

[23] Wu, X., Kan, H., & Kurths, J. (2015). A New Color Image Encryption Scheme Based on DNA Sequences and Multiple Improved 1D Chaotic Maps. Applied Soft Computing, 37, 24–39.

[24] Mondal, B., Singh, S., & Kumar, P. (2019). A secure image encryption scheme based on cellular automata and chaotic skew tent map. Journal of Information Security and Applications, 45, 117–130.

[25] Leo, M., & Kumar, R. (2021). Chaos-based secure image encryption techniques and applications. Journal of Information Security Applications.

A Diffusion-Based Generative AI Framework: An Exterior House Design from Textual Descriptions

Muhammad Amirul Akmal bin Ajusin[1], Noor Hasimah Ibrahim Teo[2]*,
Rosniza Roslan[3], Raseeda Hamzah[4], Anita Ahmad Kasim[5]
School of Computing-Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Melaka, Malaysia[1, 2, 3, 4]
Department of Information Engineering-Faculty of Engineering, Universitas Tadulako, Central Sulawesi, Indonesia[5]

***Abstract*—The architectural design process is often iterative, time-consuming, and heavily dependent on effective communication between clients and professionals. Existing design tools, such as Computer-Aided Design (CAD) systems, require technical expertise, limiting accessibility for non-professional users. This study proposes a generative artificial intelligence framework for exterior house design using a diffusion-based text-to-image model. The proposed approach integrates Stable Diffusion for image generation with a vision-language model (BLIP) to enhance semantic alignment between textual descriptions and generated outputs. In addition, an interactive refinement mechanism based on image inpainting is incorporated to allow localized modification of design elements. The system is trained on a dataset of exterior house images and evaluated using quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), as well as usability assessment. Experimental results demonstrate that the proposed framework is capable of generating semantically relevant and visually coherent architectural designs, while improving accessibility and reducing the time required for design iteration. The findings highlight the potential of generative AI as an effective tool for supporting user-centric architectural visualization and design exploration.**

***Keywords*—Generative artificial intelligence; stable diffusion; text-to-image generation; architectural design; image inpainting**

I. Introduction

Architecture plays a fundamental role in shaping human living environments by integrating functional, cultural, and aesthetic elements. The design of residential houses, particularly exterior design, requires careful consideration of architectural components such as roof structure, wall materials, windows, doors, and color schemes. Traditionally, the design process involves iterative communication between clients and architects, which can be time-consuming and inefficient due to repeated revisions and difficulty in visualizing design concepts at early stages.

Despite the availability of digital tools such as Computer-Aided Design (CAD), these systems often require significant technical expertise, limiting accessibility for non-professional users. As a result, clients face challenges in expressing their preferences effectively, while architects encounter difficulties in interpreting and translating user requirements into visual designs. This limitation highlights the need for intelligent systems that can bridge the gap between user intent and design visualization.

Recent advancements in Generative Artificial Intelligence (AI) have introduced new possibilities for automating creative tasks, including image generation and design synthesis. In particular, text-to-image generation has emerged as a promising approach that enables the transformation of natural language descriptions into visual outputs, combining techniques from natural language processing and computer vision [1], [2]. Early work in this area primarily utilized Generative Adversarial Networks (GANs), which consist of a generator and discriminator trained in an adversarial manner to produce realistic images [3]-[5]. Variants such as Conditional GAN (cGAN) [6], Deep Convolutional GAN (DCGAN) [7], and AttnGAN [8] have demonstrated improved control and semantic alignment in generated images.

Further developments introduced specialized GAN-based models for architectural and design applications. For instance, HouseGAN enables graph-constrained house layout generation [9], while Roof-GAN focuses on generating realistic roof geometries [10]. Other models, such as Pix2Pix, have been applied for image-to-image translation tasks, including architectural layout generation [11], [12]. However, GAN-based approaches often suffer from training instability, mode collapse, and limited diversity in generated outputs [13].

To address these limitations, diffusion models have gained significant attention as a more stable and robust alternative for generative tasks. Diffusion models generate images by progressively denoising random noise through a learned reverse process, allowing them to capture complex data distributions effectively [14], [15]. Denoising Diffusion Probabilistic Models (DDPM) have demonstrated superior performance in terms of image quality and diversity compared to GAN-based methods [16]. More recently, Stable Diffusion, a latent diffusion model, has enabled efficient high-resolution image generation with reduced computational cost, making it suitable for real-world applications [17].

In parallel, vision-language models have been developed to enhance the relationship between textual and visual representations. Models such as Bootstrapping Language-Image Pre-training (BLIP) leverage transformer-based architectures to generate context-aware captions and improve semantic alignment between images and text [18], [19]. These models play a crucial role in improving the accuracy and interpretability of text-to-image generation systems.

*Corresponding author.

Another important advancement in generative AI is image inpainting, which enables localized modification of images by reconstructing missing or masked regions. Modern inpainting techniques, particularly those based on diffusion models, allow users to refine specific parts of generated images while preserving global consistency [20], [21]. This capability is especially useful in architectural design, where users may wish to iteratively adjust design elements such as windows, doors, or color schemes.

Despite these technological advancements, existing research primarily focuses on model development rather than user-centric applications. Many systems lack intuitive interfaces that allow non-expert users to interact with generative models effectively. Furthermore, the integration of text-to-image generation, vision-language alignment, and interactive editing within a single architectural design framework remains underexplored.

Therefore, this study proposes a diffusion-based generative AI framework for exterior house design that integrates Stable Diffusion for image generation, BLIP for semantic alignment, and an inpainting module for interactive refinement. The proposed system is designed to be accessible to non-professional users through a chatbot-based interface, enabling intuitive and efficient design exploration. The effectiveness of the proposed approach is evaluated using quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), as well as usability assessment.

II. Methodology

A. System Design

This study adopted a Software Development Life Cycle (SDLC) perspective and implemented a modified waterfall model comprising requirements analysis, system design, implementation, testing, and documentation. In the adapted process, the deployment phase was omitted, and documentation replaced maintenance due to time and resource constraints. The overall research flow starts with image acquisition, followed by preprocessing, BLIP-based caption generation, Stable Diffusion model training, system integration, and evaluation.

The modified waterfall model was selected because the project required a structured sequence of development activities while still allowing limited feedback between phases. This was suitable for a system that combines dataset preparation, model fine-tuning, application integration, and user-based evaluation.

B. System Architecture

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. The proposed generative AI framework integrates multiple components, including data preprocessing, caption generation, model training, and user interaction, to enable automated exterior house design from textual descriptions. The system is designed as a multi-layer architecture consisting of data, application, and interaction layers.

The data and model layer includes dataset preparation, BLIP-based caption generation, and Stable Diffusion model fine-tuning. The application layer consists of a Flask-based backend that handles text-to-image generation and inpainting processes. The interaction layer is implemented using a React-based frontend that provides a chatbot interface for user input, image visualization, and design refinement.

The workflow begins when a user submits a text prompt, which is processed by the Stable Diffusion model to generate an initial house design. Users can further refine the generated output through an inpainting module that enables localized modifications. The system supports iterative interaction, allowing users to explore multiple design variations efficiently. The complete architecture and workflow of the proposed system are illustrated in Fig. 1.

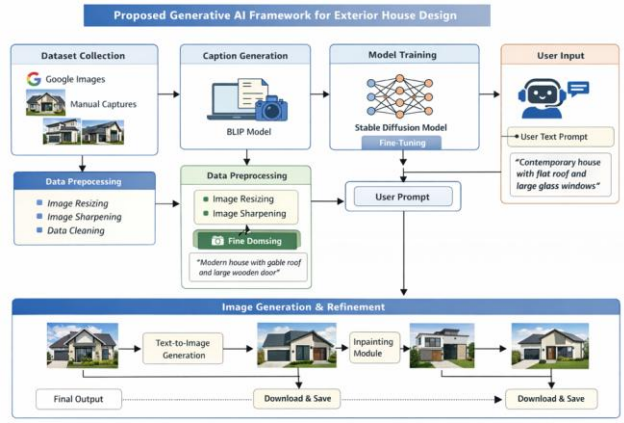


Fig. 1. Proposed generative AI framework for exterior house design.

This integrated architecture ensures seamless interaction between model components and the user interface, enabling an efficient and user-centric design workflow.

C. Dataset Collection

The dataset used in this study consists of 1,100 exterior house images, collected to represent a wide range of architectural styles and design elements relevant to residential buildings. The images were primarily obtained from publicly accessible online sources, including Google Images, using the Imageye browser extension, as well as manually captured images.

To ensure diversity, various search queries such as *"exterior house design"*, *"modern house"*, *"house with hip roof"*, and *"house with wooden door"* were used during the data collection process. This approach enabled the inclusion of different house styles, including modern, traditional, and contemporary designs.

The collected dataset focuses on front-view exterior house images and captures key architectural elements such as roof style, window type, and door design (Table I). Based on the dataset distribution, four main roof types were identified: hip roof (240 images), gable roof (275 images), flat roof (465 images), and sloped roof (120 images).

For window types, the dataset includes sliding windows (385 images), casement windows (372 images), and hung windows (343 images). In terms of door styles, wooden doors (498 images), sliding doors (397 images), and French doors (205 images) are represented.

TABLE I. DATASET STATISTICS BY ARCHITECTURAL FEATURES

Feature Category	Feature Type	Number of Images
Roof Style	Hip Roof	240
	Gable Roof	275
	Flat Roof	465
	Sloped Roof	120
Window Type	Sliding Window	385
	Casement Window	372
	Hung Window	343
Door Style	Wooden Door	498
	Sliding Door	397
	French Door	205
Total Images		**1100**

The dataset was carefully curated to remove irrelevant, duplicate, and low-quality images, ensuring that only clear and consistent exterior house images were retained. This filtering process improves the reliability of the dataset and enhances the model's ability to learn meaningful visual patterns.

Overall, the dataset provides a diverse yet structured representation of architectural features, which is essential for training a text-to-image generation model capable of producing realistic and semantically accurate exterior house designs.

D. Data Preprocessing

Data preprocessing is a crucial step to ensure that the collected dataset is suitable for training deep learning models. In this study, both image and text data undergo preprocessing to standardize input formats, enhance feature quality, and prepare structured image-caption pairs for model training.

1) Image resizing: All collected images were resized to a uniform resolution of 512 × 512 pixels to ensure consistency across the dataset. Standardizing image dimensions is essential for efficient model training, as it reduces computational complexity while preserving important architectural details such as edges, textures, and structural components.

2) Image enhancement: After resizing, image sharpening techniques were applied to enhance visual clarity and emphasize key architectural features, including roof edges, window frames, and door structures. This step improves feature distinguishability, allowing the model to better capture structural patterns during training.

3) Data cleaning: The dataset was manually cleaned to remove irrelevant, duplicate, and low-quality images. Images that did not represent exterior house views, such as interior scenes or incomplete structures, were excluded. This process ensures that the dataset remains consistent and reduces noise that may negatively affect model performance.

4) Caption generation using BLIP: To enable text-to-image learning, each image must be paired with a corresponding textual description. In this study, captions were generated using the Bootstrapping Language-Image Pre-training (BLIP) model. The BLIP model was fine-tuned on the dataset to produce descriptive captions that capture key architectural elements such as roof type, window style, door design, and wall characteristics.

The generated captions serve as textual representations of the visual features, allowing the model to learn the relationship between natural language input and architectural design elements.

5) Dataset structuring: The final dataset was organized into structured image-text pairs in the form:

(Image, Caption)

This structured format is essential for training the Stable Diffusion model, as it enables supervised learning of the mapping between textual prompts and corresponding visual outputs. An example of an image-caption pair from the dataset is shown in Fig. 2. The caption describes key architectural elements, including roof type, window style, door design, and wall color, which are used as conditioning inputs during model training.



Fig. 2. Example of image-caption pair used in the dataset: "Modern house with flat roof, large sliding windows, white wall, and wooden front door."

The example illustrates how textual descriptions are aligned with visual features, allowing the model to learn meaningful relationships between language and architectural design elements. This structured pairing plays a crucial role in improving the semantic accuracy of the generated images.

6) Data preparation for model training: The processed dataset was prepared for training by ensuring compatibility with the Stable Diffusion framework. Image-caption pairs were formatted and stored in a suitable structure for fine-tuning using the DreamBooth method. Proper data preparation improves training stability and contributes to the generation of high-quality and semantically accurate images.

E. System Design

This section describes the development of the core models used in the proposed framework, namely the BLIP model for image captioning and the Stable Diffusion model for text-to-image generation. The integration of these models enables the system to learn the relationship between textual descriptions and architectural visual features.

1) BLIP model for caption generation: The Bootstrapping Language-Image Pre-training (BLIP) model is employed to generate descriptive captions for the exterior house images. Initially, a pre-trained BLIP model is loaded and fine-tuned using the prepared dataset to adapt it to the domain of architectural design.

During training, the BLIP processor converts both images and associated textual descriptions into numerical representations, allowing the model to learn semantic relationships between visual features and language. The fine-tuning process improves the model's ability to generate captions that accurately describe architectural elements such as roof style, window type, door design, and wall characteristics.

Once trained, the BLIP model is used to automatically generate captions for all images in the dataset. These captions form structured image-text pairs that serve as input for training the text-to-image generation model.

2) Stable diffusion model for text-to-image generation: The Stable Diffusion model is used as the primary generative model for producing exterior house designs from textual prompts. Stable Diffusion operates by transforming random noise into meaningful images through an iterative denoising process guided by text embeddings.

In this study, a pre-trained Stable Diffusion model is fine-tuned using the prepared image-caption dataset. The fine-tuning process is performed using the DreamBooth approach, which enables the model to adapt to domain-specific features while preserving general image generation capabilities.

During training, textual captions are encoded into embeddings and used to condition the denoising process. A U-Net architecture progressively removes noise from a latent representation, while a decoder converts the final latent representation into a high-resolution image. This process allows the model to generate visually coherent images that correspond to user-defined prompts.

3) Model integration: The BLIP and Stable Diffusion models are integrated to form a unified generative pipeline. The BLIP model provides high-quality captions that enhance semantic alignment, while the Stable Diffusion model generates images based on these textual descriptions.

This integration improves the overall performance of the system by ensuring that generated images are both visually realistic and semantically relevant. The combined framework enables effective mapping between user input and generated architectural designs.

4) Model deployment: After fine-tuning, the Stable Diffusion model is exported into the Hugging Face Diffusers format to support efficient inference. The trained model is then deployed within a Flask-based backend, which processes user prompts and returns generated images in real time.

This deployment setup ensures scalability and enables seamless interaction between the frontend interface and the generative model.

F. Proposed Framework Algorithm

The overall workflow of the proposed generative AI framework is formalized in Algorithm 1, which integrates data preprocessing, caption generation, model training, text-to-image generation, and interactive refinement.

Algorithm 1:

```
Input:
   Raw image dataset I_raw
   User prompt p
Output:
   Generated or refined exterior house image g

// DATA PREPARATION PHASE
Collect exterior house images I_raw from online sources
For each image image_i in I_raw do
    Resize image_i to 512 × 512
    Apply sharpening to enhance architectural features
    Store processed image_i in dataset I_clean
End For

// CAPTION GENERATION (BLIP)
Load pre-trained BLIP model
Fine-tune BLIP using I_clean
For each image image_i in I_clean do
    Generate caption_i using BLIP
    Store (image_i, caption_i) in dataset D
End For

// MODEL TRAINING (STABLE DIFFUSION)
Load pre-trained Stable Diffusion model M_sd
Fine-tune M_sd using dataset D (image-caption pairs)
Save trained model

// INFERENCE PHASE (TEXT-TO-IMAGE)
Receive user prompt p from interface
Encode p into text embeddings
Initialize latent vector z with random noise
For t = T down to 1 do
    Predict noise using U-Net conditioned on p
    Denoise z iteratively
End For
Decode final latent z into generated image g

// OPTIONAL INPAINTING (REFINEMENT)
If user selects region for editing then
    Receive mask m and refinement prompt p_r
    Preserve unmasked regions of g
    Apply diffusion-based inpainting on masked region
    Generate refined image g'
    Set g ← g'
End If

// OUTPUT
Display generated image g to user
Allow download or further refinement

Return g
```

Algorithm 1 illustrates the complete pipeline from dataset preparation to user interaction, highlighting the integration of diffusion-based generation and vision-language alignment within a unified framework.

III. Result and Discussion

This section presents the evaluation results of the proposed generative AI framework for exterior house design. The system is assessed based on functionality, semantic accuracy, image quality, and usability. Quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), are used to evaluate model performance, while user-based testing is conducted to assess system usability.

A. Functionality Testing

Functionality testing was conducted to verify that the proposed system operates correctly in supporting the main user tasks. The evaluation focused on validating the core features of the system, including text-to-image generation, image refinement through inpainting, user interface interaction, and image output management.

The following key features were tested during the functionality evaluation:

- Text prompt input through the chatbot interface.
- Text-to-image generation using the Stable Diffusion model.
- Display of generated images.
- Image variation generation.
- Inpainting-based image refinement.
- Masking tools (brush, eraser, and selection tools).
- Navigation between Design Generator and Inpainting Studio.
- Image download functionality.

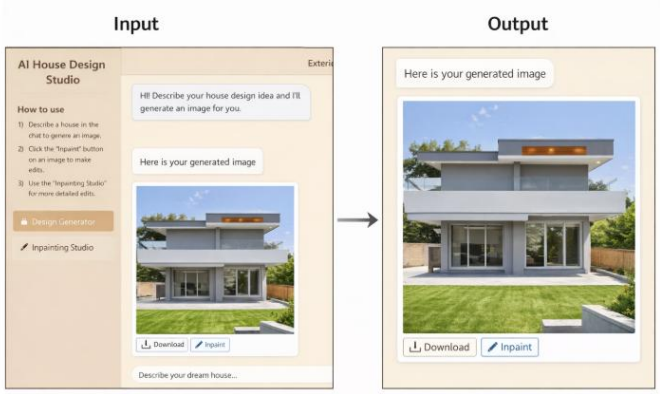


Fig. 3. Example of text-to-image generation using a prompt.

For the text-to-image generation module, the system successfully accepted natural language prompts entered through the chatbot interface and produced corresponding exterior house design images. As illustrated in Fig. 3, the generated result reflects the architectural characteristics described by the user, indicating that the generation pipeline is able to transform textual input into a coherent visual output. This demonstrates that the model and interface are properly integrated and capable of supporting intuitive design exploration.

The system also performed reliably in supporting image refinement through the inpainting module. Users were able to select a specific region of the generated image, apply a mask, and submit an edit prompt to modify the selected area. As shown in Fig. 4, the system successfully updated the targeted region while preserving the overall layout and visual consistency of the house design. This confirms the effectiveness of the inpainting function for localized architectural refinement.



Fig. 4. Inpainting-based refinement of generated house design.

In addition, supporting features such as switching between modules, selecting generated variations, and downloading images were executed without errors. The masking tools (e.g., brush and eraser) also functioned correctly, allowing users to interactively define regions for modification.

Overall, the functionality testing results indicate that all major system components operate as expected. The successful execution of generation, refinement, and interaction features demonstrates that the proposed framework is fully functional and suitable for assisting users in exterior house design visualization.

B. Semantic Accuracy Evaluation (CLIP Score)

The semantic alignment between generated images and their corresponding textual descriptions was evaluated using the CLIP Score. This metric measures the similarity between image and text embeddings, with higher values indicating better semantic consistency.

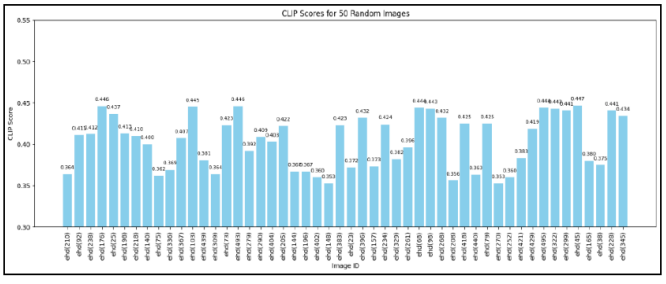


Fig. 5. CLIP score visualization.

The BLIP model achieved an average CLIP Score of 0.42, indicating moderate alignment between generated captions and visual features as shown in Fig. 5. This result suggests that the model is capable of capturing relevant architectural elements

from textual descriptions, although some discrepancies may still occur in complex scenarios.

The moderate CLIP score can be attributed to the limited dataset size and variability in architectural styles, which may affect the model's ability to generalize across diverse design inputs.

C. Image Quality Evaluation (FID Score)

The visual quality of the generated images was evaluated using the Fréchet Inception Distance (FID), which measures the similarity between the distributions of generated images and real images in terms of feature representations. Lower FID values indicate better image quality and higher similarity to real-world data.

The Stable Diffusion model achieved an FID score of 82.4, indicating that the generated images exhibit recognizable architectural structures but still lack high levels of photorealism and fine-grained detail. Despite the relatively high FID score, qualitative observation of the generated images shows that the model is capable of producing coherent house designs with identifiable architectural elements such as roof structures, windows, doors, and wall compositions. This suggests that the model successfully captures the general layout and structural characteristics of exterior house designs, even if certain visual refinements are not fully achieved.

The obtained FID score can be attributed to several factors. First, the dataset size of 1,100 images is relatively limited compared to large-scale datasets typically used in state-of-the-art generative models. This restricts the model's ability to learn diverse architectural patterns and may lead to less realistic outputs. Second, the dataset exhibits some imbalance in feature distribution, particularly with a higher proportion of flat roof and wooden door designs, which may bias the generated results toward more common styles.

In addition, the complexity of architectural features poses a challenge for generative models. Unlike general object generation, house design involves structured geometry, symmetry, and spatial consistency, which require the model to capture both global layout and local details simultaneously. The current model demonstrates competence in generating overall structure but shows limitations in rendering fine textures, lighting consistency, and detailed architectural features.

Furthermore, the fine-tuning process using the DreamBooth approach, while effective for adapting the model to a specific domain, may not fully optimize the model for high-fidelity image generation when trained on a relatively small dataset. This highlights the trade-off between domain specialization and image realism.

When compared to state-of-the-art diffusion models trained on large-scale datasets, the obtained FID score is higher; however, within the scope of this study, the generated images remain sufficiently realistic for early-stage architectural visualization and concept exploration. The primary objective of the system is to support idea generation rather than produce final construction-ready designs, and thus, the achieved image quality is considered acceptable for its intended application.

To improve image quality in future work, several enhancements can be considered. These include increasing the dataset size and diversity, applying data augmentation techniques, incorporating higher-resolution training strategies, and fine-tuning more advanced diffusion architectures. Additionally, integrating post-processing techniques or hybrid models could further enhance visual realism.

Overall, the FID evaluation indicates that while there is room for improvement in image realism, the proposed system successfully generates meaningful and visually coherent architectural designs, demonstrating the feasibility of using diffusion-based generative models for exterior house design applications.

D. Usability Testing

Usability testing was conducted to evaluate the effectiveness, efficiency, and user satisfaction of the proposed system. The assessment was performed using the System Usability Scale (SUS), a widely adopted method for measuring perceived usability of interactive systems. The SUS questionnaire consists of ten items rated on a five-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." The scoring method involves subtracting 1 from the scores of odd-numbered questions and subtracting the scores of even-numbered questions from 5, followed by multiplying the total score by 2.5 to obtain a final value between 0 and 100.

A total of 30 respondents participated in the usability evaluation. The results, as illustrated in Fig. 6, show that most respondents achieved SUS scores within the range of 80 to 95. The overall average SUS score obtained in this study is 86, indicating a high level of usability and user satisfaction.

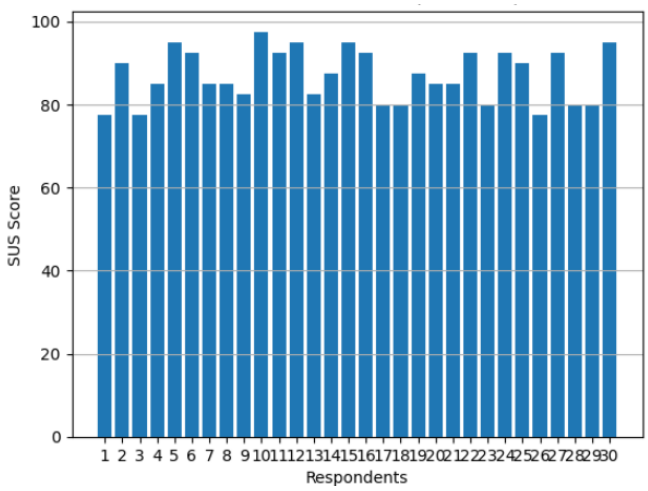


Fig. 6. SUS score distribution.

According to standard SUS interpretation guidelines, a score above 68 is considered above average, while scores above 80 are classified as excellent. Therefore, the obtained average score of 86 places the proposed system within the "excellent" usability category, suggesting that users found the system easy to use and well-designed. The high usability score can be attributed to several design factors. First, the chatbot-based interface simplifies user interaction by allowing natural language input, eliminating the need for technical knowledge in architectural

design tools. Second, the system provides immediate visual feedback through generated images, which enhances user engagement and supports rapid design exploration. Third, the inclusion of interactive features such as image variation and inpainting allows users to refine designs intuitively, contributing to a more flexible and user-centered experience.

Furthermore, the relatively consistent SUS scores across respondents indicate that the system provides a stable and reliable user experience. The absence of significantly low scores suggests that users did not encounter major usability issues during interaction. However, some limitations were observed. A few respondents reported slightly lower scores, which may be attributed to the need for clearer guidance when using advanced features such as inpainting or masking tools. This indicates that additional user assistance, such as tooltips or tutorials, could further improve the overall usability of the system.

Overall, the usability evaluation demonstrates that the proposed framework successfully meets user expectations in terms of ease of use, functionality, and interaction. The results highlight the potential of the system to serve as an accessible tool for both non-expert users and professionals in architectural design

E. Discussion

The experimental results demonstrate that the proposed generative AI framework is capable of generating exterior house designs from textual descriptions with reasonable semantic alignment and visual coherence. The integration of BLIP and Stable Diffusion enables effective mapping between text and image, while the inpainting module provides flexibility for iterative design refinement.

The system offers several advantages over traditional design approaches. First, it reduces the time required for design iteration by providing instant visual feedback. Second, it improves communication between clients and architects by enabling users to visualize their ideas without requiring technical expertise. Third, it introduces an interactive and user-friendly design process through the chatbot interface.

However, several limitations were identified. The moderate CLIP score indicates that semantic alignment can be further improved, particularly for complex or ambiguous prompts. Additionally, the relatively high FID score suggests that image realism is limited by the dataset size and diversity. Expanding the dataset and incorporating more advanced training techniques could improve model performance.

Overall, the results highlight the potential of generative AI as a practical tool for architectural visualization, particularly in early-stage design and concept exploration.

IV. CONCLUSION

This study presented a diffusion-based generative AI framework for exterior house design that enables the transformation of textual descriptions into architectural visualizations. The proposed system integrates Stable Diffusion for text-to-image generation with a BLIP-based captioning model to improve semantic alignment, along with an inpainting module for interactive design refinement. The experimental results demonstrate that the proposed framework is capable of generating visually coherent and semantically relevant exterior house designs. The system achieved a CLIP Score of 0.42, indicating moderate alignment between textual descriptions and generated images, and an FID score of 82.4, suggesting acceptable image quality for early-stage design visualization. In addition, usability evaluation shows that the system is intuitive and accessible, allowing non-professional users to explore design ideas effectively.

The main contributions of this study include the development of an end-to-end generative AI framework for architectural design, the integration of diffusion models with vision-language techniques, and the introduction of an interactive refinement mechanism using inpainting. These contributions address limitations in traditional design workflows by reducing dependency on technical expertise and improving communication between users and designers. Despite these contributions, several limitations remain. The relatively high FID score indicates that the generated images can be further improved in terms of realism and fine details. Additionally, the dataset size and diversity are limited, which may affect the model's ability to generalize across complex architectural styles.

Future work will focus on expanding the dataset with more diverse architectural designs, improving model performance through advanced training strategies, and extending the system to support interior design and 3D architectural generation. Furthermore, integrating real-time user feedback and deploying the system as a web or mobile application could enhance its practical applicability.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Teknologi MARA for the financial and matching grant administrative support. This research was funded by Universiti Teknologi MARA (UiTM) provided through Geran Sepadan TEJA (GST 2025/1-2).

REFERENCES

[1] L. Sudha, K. B. Aruna, V. Sureka, M. Niveditha, and S. Prema, "Semantic image synthesis from text: Current trends and future horizons in text-to-image generation," EAI Endorsed Transactions on Internet of Things, vol. 11, 2024.

[2] Z. Lin et al., "Evaluating text-to-visual generation with image-to-text generation," in Proc. European Conf. Computer Vision (ECCV), Cham, Switzerland: Springer, 2024, pp. 366–384.

[3] A. Dash, J. Ye, and G. Wang, "A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: From medical to remote sensing," IEEE Access, vol. 12, pp. 18330–18357, 2023.

[4] T. Chakraborty et al., "Ten years of generative adversarial nets (GANs): A survey of the state-of-the-art," Machine Learning: Science and Technology, vol. 5, no. 1, p. 011001, 2024.

[5] V. L. T. de Souza et al., "A review on generative adversarial networks for image generation," Computers & Graphics, vol. 114, pp. 13–25, 2023.

[6] S. S. Mohammed and H. G. Clarke, "Conditional image-to-image translation generative adversarial network (cGAN) for fabric defect data augmentation," Neural Computing and Applications, vol. 36, no. 32, pp. 20231–20244, 2024.

[7] J. Jenkins and K. Roy, "Exploring deep convolutional generative adversarial networks (DCGAN) in biometric systems: A survey study," Discover Artificial Intelligence, vol. 4, no. 1, p. 42, 2024.

[8] R. Gopalakrishnan, N. Sambagni, and P. V. Sudeep, "An improved AttnGAN model for text-to-image synthesis," in Proc. Int. Conf.

Computer Vision and Image Processing, Cham, Switzerland: Springer, 2023, pp. 139–151.

[9] N. Nauata, K.-H. Chang, C.-Y. Cheng, G. Mori, and Y. Furukawa, "House-GAN: Relational generative adversarial networks for graph-constrained house layout generation," in Proc. ECCV, 2020, pp. 162–177.

[10] H. Tang et al., "Graph transformer GANs for graph-constrained house generation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2173–2182.

[11] Z. Li et al., "Mapping new realities: Ground truth image creation with pix2pix image-to-image translation," arXiv preprint arXiv:2404.19265, 2024.

[12] X. Zhao, H. Yu, and H. Bian, "Image to image translation based on differential image Pix2Pix model," Computers, Materials & Continua, vol. 77, no. 1, 2023.

[13] M. Cobbinah et al., "Diversity in stable GANs: A systematic review of mode collapse mitigation strategies," Engineering Reports, vol. 7, no. 6, e70209, 2025.

[14] J. Liu et al., "Residual denoising diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2773–2783.

[15] A. K. Bhunia et al., "Person image synthesis via denoising diffusion model," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5968–5976.

[16] G. Müller-Franzes et al., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," Scientific Reports, vol. 13, no. 1, p. 12098, 2023.

[17] D. Podell et al., "SDXL: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.

[18] P. Panchal et al., "Deep learning–driven image captioning: Progress through transformers and large language models," PLoS ONE, vol. 21, no. 3, p. e0345012, 2026.

[19] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image–text alignment for remote sensing image captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–12, 2024.

[20] C. Corneanu, R. Gadde, and A. M. Martinez, "LatentPaint: Image inpainting in latent space with diffusion models," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2024, pp. 4334–4343.

[21] A. Lugmayr et al., "RePaint: Inpainting using denoising diffusion probabilistic models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11461–11471.

Integration of Random Forest and Hough Transform for Cancer Classification Using Microarray Gene Expression Data

Hibah Alatawi*, Hechmi Shili
Department of Computer Sciences, University of Tabuk, Saudi Arabia

***Abstract*—Cancer classification poses a significant challenge owing to the intricate nature and diversity of the disease. This study introduces a novel methodology for cancer classification leveraging microarray gene expression data. The proposed approach integrates Random Forest (RF) and Hough Transform (HT), where RF performs feature selection and classification, and HT identifies sub-biclusters that are merged into larger clusters using a hypergraph model. Evaluation on multiple cancer datasets demonstrates that the hybrid approach improves classification accuracy compared to standalone RF or HT while capturing meaningful gene expression patterns.**

***Keywords*—Cancer classification; microarray gene expression; random forest; Hough transform; feature selection; hypergraph model**

I. Introduction

Microarray data analysis plays a pivotal role in understanding complex biological systems and identifying biomarkers associated with various diseases, including cancer [1,10]. Microarray analysis is employed in gene expression profiling, where it typically generates a substantial amount of data [31,32]. Biclustering algorithms have emerged as a powerful tool in the analysis of microarray gene expression data, allowing for the simultaneous classification of both genes and experimental conditions. Unlike standard clustering techniques, which focus on grouping either genes or samples, biclustering methods aim to identify sub-matrices within the gene expression data matrix where a subset of genes exhibit a consistent pattern across a subset of experimental conditions [1]. In recent years, the classification of microarray datasets, often referred to as large-scale biological data analysis, has become a prominent and compelling area of research. The application of microarray technology represents a critical advancement in molecular biology, particularly for the detection of cancer [37]. In Addition, biclustering, a powerful data mining technique, facilitates the simultaneous clustering of rows and columns in large datasets, uncovering hidden patterns and structures. However, traditional biclustering methods often struggle with the high dimensionality and noise inherent in microarray datasets [2]. Microarray technology has become a powerful tool for investigating complex biological processes and diseases, enabling the representation of gene expression data as large matrices with genes as rows and experimental conditions as columns [3]. While various statistical techniques have been applied to analyze these gene expression matrices, biclustering has emerged as a particularly effective method, as it can uncover meaningful patterns by identifying subsets of genes that exhibit consistent expression profiles across subsets of experimental conditions. This ability to capture the inherent complexity of gene regulation, where the expression of a gene can be influenced by multiple factors, is a key advantage of biclustering over traditional clustering approaches, which focus on grouping genes or samples independently. The insights gained from biclustering analysis can provide valuable perspectives on the intricate mechanisms governing biological processes and disease states [4].

Despite these advances, existing biclustering and classification approaches still face limitations in handling the high dimensionality, noise sensitivity, and the difficulty of linking biclustering outputs directly to robust classification performance. This gap motivates the need for an integrated and more resilient framework for microarray data analysis.

An automated, computer-assisted medical diagnosis system that integrates cutting-edge medical methodologies with advanced machine learning algorithms represents a critical multidisciplinary technology. It enables accurate and noninvasive diagnosis of various diseases, including breast cancer [6,32]. The classification of microarray data presents significant challenges due to its complexity. The bioinformatics community employs a variety of approaches to diagnose and classify this data using machine learning systems. Given that each microarray sample contains thousands of genes for analysis, feature elimination methods have become widely utilized and adopted to manage the high dimensionality and enhance the accuracy of classification outcomes [5,39]. The abundance of gene expression datasets offers a valuable opportunity to computationally identify condition-specific functional gene modules (FGMs). These FGMs are characterized by highly organized expression patterns within specific sets of genes. Smart healthcare involves a wide range of operatives, including physicians, nurses, hospitals, and research organizations [30].

This study aims to develop an integrated framework that combines geometric pattern detection and ensemble learning to improve both biclustering quality and classification accuracy in microarray gene expression data.

In this study, we propose a novel methodology that addresses these challenges by integrating the Hough Transform and Random Forest algorithm for biclustering and classification tasks. The Hough Transform, applied in column pair-spaces, enables the identification of sub-biclusters within microarray

*Corresponding author.

datasets, while hyper-graph partitioning techniques refine and optimize these sub-biclusters. Subsequently, the optimized biclusters are seamlessly integrated as inputs into a Random Forest classifier, enhancing the classification process. By combining these techniques, our integrated framework offers a comprehensive solution for biclustering and classification tasks in Microarray data analysis. Leveraging the complementary strengths of the Hough Transform and Random Forest algorithm, our approach provides a holistic and efficient method for uncovering meaningful patterns and biomarkers in complex biological datasets.

The main contributions of this work are summarized as follows:

- Proposing a novel integration between the Hough Transform and Random Forest for simultaneous biclustering and classification.
- Introducing a hyper-graph partitioning refinement step to optimize detected sub-biclusters.
- Providing a robust framework that links biclustering outputs directly to classification performance.
- Demonstrating the effectiveness of the proposed approach on real-world microarray cancer datasets.

The integration of biclustering and classification offers several advantages. First, it enables the simultaneous identification of coherent gene expression patterns and their classification into distinct biological states or phenotypes. Second, by incorporating the Hough Transform, our framework provides robustness to noise and captures complex patterns inherent in Microarray gene expression data. Third, the Random Forest classifier enhances the accuracy and robustness of classification by leveraging the collective decision of multiple decision trees. The proposed framework holds great promise for advancing our understanding of gene expression regulation and its role in disease mechanisms. By providing a holistic approach to Microarray gene expression data analysis, it offers researchers a powerful tool for uncovering meaningful insights and identifying potential biomarkers associated with diseases such as cancer.

Unlike existing studies that treat biclustering and classification as separate processes, the proposed approach establishes a direct and optimized connection between both tasks, highlighting the novelty of this work compared to recent related studies.

In the subsequent sections of this study, we describe the methodology in detail, present experimental results on real-world datasets, and discuss the implications of our findings for biomedical research and clinical applications.

II. Related Work

The field of gene expression analysis has witnessed substantial advancements, particularly in the development and application of biclustering techniques. Biclustering, a method that allows for the simultaneous classification of gene expression profiles, has proven to be effective in identifying patterns of gene expression under diverse conditions. This literature review examines the evolution of biclustering methods and their application in microarray gene expression data analysis. Recent studies (Table I) have introduced innovative frameworks and algorithms that enhance the classification of gene expression profiles, including the use of Biclustering-based classifiers, evolutionary approaches, and ensemble learning techniques. Additionally, the integration of machine learning and big data approaches has further refined cancer diagnosis and classification, demonstrating significant improvements in accuracy and robustness. This section also highlights the potential of combining traditional statistical methods with advanced computational techniques, such as the Hough Transform, to improve the analysis and interpretation of complex biological data. Through a comprehensive examination of these advancements, the review underscores the critical role of biclustering and related methodologies in advancing our understanding of gene expression patterns and improving cancer classification and diagnosis.

A. Advances in Biclustering and Machine Learning for Gene Expression and Cancer Diagnosis

Biclustering, a method for simultaneous classification of gene expression profiles, has been shown to be effective in identifying gene expression patterns under various conditions [1]. This method has been further enhanced by the development of a biclustering-based classification framework, which utilizes homogeneously expressed genes in biclusters to construct a classifier for sample class membership prediction Malhotra et al. [2]. An evolutionary approach to obtaining high-quality biclusters of highly-correlated genes has also been proposed, demonstrating competitive performance with state-of-the-art algorithms Ayadi et al. [3]. Additionally, the application of the bagging approach to improve the performance of biclustering methods has been successful in both synthetic and real datasets Hanczar, Ayadi et al. [4]. These studies collectively highlight the potential of biclustering and its various applications in microarray gene expression data analysis. The proposed framework of Biswal et al. [7] undergoes testing on a diverse set of datasets, including five distinct microarray datasets and one dataset consisting of single-cell RNA sequencing (scRNA-seq) data. In Albalawi et al. [26], the Convolutional Neural Network (CNN) classifier for diagnosing breast cancer utilizing MIAS (Mammographic Image Analysis Society) dataset. CNN established as an efficient class of methods for image recognition problems. The research of Almutairi, S. et al. [27], develops a big data-driven breast cancer classification model using Deep Reinforcement Learning (DRL), achieving high accuracy across three datasets (WBCD: 98.90%, WDBC: 99.02%, WPBC: 98.88%). The model incorporates the Gorilla Troops Optimization (GTO) algorithm for feature selection and Deep Q Learning (DQL) for classification, outperforming traditional methods such as RBF-ELB, PSO-MLP, and GA-MLP. The study demonstrates the potential of integrating big data and advanced machine learning techniques to improve breast cancer diagnosis accuracy. The study of Alatawi, Y. M. et al. [28] assessed early performance indicators for breast cancer screening at King Abdulaziz University Hospital, Saudi Arabia, analyzing data from 1,911 women who participated in the screening program between 2012 and 2019. The findings revealed that 19.9% of women were recalled for further evaluation, with 18.9% undergoing biopsy. Screen-detected

cancer was found in 1.6% of participants, while 0.7% were diagnosed with breast cancer.

Balaha and Hassan [33] proposed a deep learning-based framework for skin cancer diagnosis by integrating transfer learning with the Sparrow Search Algorithm (SSA) to optimize feature extraction and classification performance. Their approach demonstrated superior results compared with several conventional deep learning architectures, achieving high accuracy, precision, recall, and F1-score in skin lesion classification. Since the Hough transform is an image analysis technique, it can potentially be integrated with other computational platforms and machine learning models to further enhance image-based diagnostic performance [34].

B. Computational Models for Microarray-Based Cancer Classification

Recent advances in biomedical data analysis have highlighted the potential of image-derived mathematical models, such as the Hough Transform (HT), in uncovering hidden structural relationships within complex biological datasets. Originally developed for geometric feature extraction in image processing, HT has been successfully adapted to identify spatial and relational patterns in biomedical contexts, including histopathological imaging, brain signal analysis, and genomic data interpretation. In particular, the HT framework enables the detection of geometric trends—such as alignments or co-expressed trajectories—among gene pairs or condition profiles. This geometric abstraction provides an effective means of representing high-dimensional biological data in lower-dimensional spaces, facilitating the discovery of regulatory modules and functional relationships that traditional statistical methods may overlook [1]. Beyond geometric modeling, several computational frameworks have explored the synergy between biclustering and feature selection to improve cancer classification using microarray gene expression data. Biclustering methods—such as the Cheng & Church algorithm, which minimizes mean squared residue to identify locally coherent expression patterns; xMotif, which searches for statistically significant submatrices enriched with co-expressed genes; and the Plaid model, which decomposes the expression matrix into overlapping layers representing distinct biological processes—have been widely employed to capture coherent gene subsets that exhibit correlated expression under specific experimental conditions [40]. Feature selection algorithms, including mutual information, ReliefF, and ensemble-based approaches, have been integrated with machine learning classifiers to reduce dimensionality and enhance interpretability. Among these, the Random Forest (RF) classifier has emerged as a preferred model due to its robustness to noise, built-in feature importance estimation, and ability to handle high-dimensional gene expression data efficiently. Recent studies have demonstrated that RF-based hybrid systems outperform conventional models in identifying diagnostic biomarkers and distinguishing between cancer subtypes, motivating the integration of RF with biclustering and structural learning in this study.

C. Advances in Gene Expression Analysis

The Hough Transform has been successfully applied in the analysis of microarray gene expression data. Zhao et al [8] and Zhao et al. [9] both developed biclustering algorithms based on the Hough Transform in the column-pair space, which were found to be robust to noise and computationally efficient. Tiňo et al. [10] further improved this approach by introducing a probabilistic model-based Hough Transform for the detection of co-expression patterns in three-color cDNA microarray data. Shafie et al. [11] extended this work by demonstrating the use of a relational space to classify microarray gene expression data, with the transformation of real-valued data to binary data leading to improved class separation. These studies collectively highlight the potential of the Hough Transform in the analysis of microarray gene expression data. Over the past three decades, advancements in cancer classification have been significant; however, a universal methodology for the identification of new cancer classes (class discovery) and the assignment of tumors to established classes (class prediction) remains elusive [12]. Golub et al. [12,16] presented a comprehensive approach to cancer classification utilizing gene expression monitoring via DNA microarrays, specifically applied to human acute leukemias as an illustrative case. The study of Umer, M. et al. [29] introduces an ensemble learning-based voting classifier that integrates logistic regression and stochastic gradient descent classifiers with deep convoluted features for the accurate detection of breast cancer. The proposed framework effectively distinguishes between malignant and benign tumors, demonstrating improved classification accuracy. The results indicate that this method outperforms existing state-of-the-art approaches. The class discovery process successfully differentiated between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without prior knowledge of these categories. Furthermore, a class predictor derived from this process accurately classified new leukemia cases. These findings validate the viability of cancer classification based exclusively on gene expression profiles and propose a generalizable strategy for the discovery and prediction of cancer classes across various cancer types, independent of pre-existing biological knowledge. A hybrid approach of PCC-BPSO/GA with multiple classifiers improves gene selection and classification accuracy in microarray datasets [17]. Differentiating malignant pleural mesothelioma (MPM) from adenocarcinoma (ADCA) of the lung is challenging with conventional methods. G. Gordon et al. [13,15] proposed a gene expression ratio-based approach for early and precise diagnosis.

This study assessed 181 tissue samples (31 MPM and 150 ADCA) and identified 15 diagnostic ratios from a training set of 32 samples. Each ratio achieved at least 90% accuracy in diagnosis, with combined ratios reaching 95% and 99% accuracy for MPM and ADCA, respectively. This method is accurate, cost-effective, and has potential applicability in other clinical diagnostic contexts.

Extensive experiments in Purbolaksono et al. [16] on 8 real cancer microarray datasets (4 diagnostic and 4 prognostic) show that the proposed classifier performed superior in both cancer diagnosis and prognosis, the latter of which was regarded as quite difficult previously. Additionally, results demonstrate that sample classification accuracy can serve as a good subjective quality measure for different types of biclusters, and hence as a tool to extrinsically evaluate the performance of various biclustering algorithms that produce those biclusters.

Microarray data classification poses significant challenges due to the high dimensionality and small sample sizes of the data.

V. Bolon et al. [18] present an extensive experimental evaluation of the most prominent feature selection algorithms and evaluation techniques, providing a detailed analysis of the findings and insights that can inform the development of effective microarray data classification approaches. Embryonal tumors of the central nervous system (CNS) represent a biologically diverse and poorly understood group of tumors. Diagnosing these tumors based solely on their morphology is controversial, as exemplified by medulloblastomas—the most common malignant brain tumors in children—whose pathogenesis and relationship to other embryonal CNS tumors remain debated. Additionally, predicting patients' responses to therapy for these tumors poses a significant challenge. Pomeroy SL et al. [19] addressed these issues by developing a classification system based on DNA microarray gene expression data from 99 patient samples. They demonstrated that medulloblastomas are molecularly distinct from other brain tumors, including primitive neuroectodermal tumors (PNETs), atypical teratoid/rhabdoid tumors (AT/RTs), and malignant gliomas. Their research provided previously unrecognized evidence that medulloblastomas derive from cerebellar granule cells through activation of the Sonic Hedgehog (SHH) pathway. Furthermore, they showed that the clinical outcomes of children with medulloblastomas are highly predictable based on the gene expression profiles of their tumors at diagnosis. Jinthanasatian et al. [20] proposed a new feature selection algorithm called FF-SVM that uses the FireFly algorithm along with an SVM classifier to classify cancer microarray gene expression data, and they found that this algorithm achieves high classification accuracy using a small number of selected genes. Salem et al. [21] present a novel methodology for classifying human cancer diseases based on gene expression profiles. This approach combines Information Gain (IG) and Deep Genetic Algorithm (DGA).

TABLE I. SUMMARY OF STUDIES ON BI-CLUSTERING TECHNIQUES IN MICROARRAY GENE EXPRESSION DATA ANALYSIS [15]

Ref	Dataset	Type classifier	Feature Selection	Accuracy	No.Genes
[16]	Leukemia [12]	BN	Ml	88.2%	NA
		NN	FCBF	99.44%	51
		Decision Table	CFS	98.6%	79
[17]	Lung Cancer [13]	NB	Ml	98.66%	NA
		NN	SFFS	97.92%	308
		NB	PCC-BPSO	98.03%	39
		RF	K-means - Relief	98.90%	NA
[18]	Prostate [14]	SVM	ReliefF	97%	50
		ELM	MIMAGA	97.12%	60
		RF	K-means -Relief	88.97%	NA
[20]	DLBCL	rule set generation	a neuro-fuzzy algorithm	83.81%	13
[21]	DLBCL	GP	IG-SGA	94.80%	110

III. PROPOSED FRAMEWORK

Microarray data analysis plays a pivotal role in understanding complex biological systems and identifying biomarkers associated with various diseases, including cancer. Biclustering, a powerful data mining technique, facilitates the simultaneous clustering of rows and columns in large datasets, uncovering hidden patterns and structures. However, traditional biclustering methods often struggle with the high dimensionality and noise inherent in microarray datasets. In this study, we propose a novel methodology that addresses these challenges by integrating the Hough Transform and Random Forest algorithm for biclustering and classification tasks. The Hough Transform, applied in column pair-spaces, enables the identification of sub-biclusters within microarray datasets, while hypergraph partitioning techniques refine and optimize these sub-biclusters. Subsequently, the optimized biclusters are seamlessly integrated as inputs into a Random Forest classifier, enhancing the classification process.

Fig. 1 summarizes the full workflow of the proposed framework, from the input microarray matrix to the final classification result. The process starts by projecting gene expression data into column-pair spaces, followed by Hough Transform–based detection of geometric patterns to extract sub-biclusters. These are encoded, grouped by shared columns, and used to build a hypergraph that is partitioned and optimized to obtain refined biclusters. Features are then derived from the optimized biclusters and fed into a Random Forest classifier for cancer prediction. The next subsections explain each stage step by step.

A. Geometric Patterns in Microarray Data Using Hough Transform in Column-Pair Spaces

Microarray gene expression data often exhibit complex geometric patterns that encode valuable information about underlying biological processes. The application of the Hough Transform (HT) in column-pair spaces offers a powerful tool for uncovering these geometric patterns and extracting meaningful insights from the data. In microarray experiments, gene expression levels are typically measured across multiple conditions (columns) for numerous genes (rows). Traditional clustering methods often overlook the spatial relationships between genes and conditions, limiting their ability to capture

the intricate geometric structures present in the data. However, by leveraging the HT in column-pair spaces, we can explore these geometric patterns more effectively. The HT is a robust technique originally developed for line detection in image processing. When applied to microarray data, it allows us to identify linear structures or patterns within the dataset, revealing relationships between genes and conditions that may not be apparent through conventional methods. In column-pair spaces, each point represents the expression levels of a gene across two conditions. By transforming the data into this space, we can detect linear patterns corresponding to coherent gene expression changes across specific conditions. These patterns may manifest as diagonal lines, indicating consistent up- or down-regulation of genes under particular conditions, or as clusters of points representing subsets of genes exhibiting similar expression profiles across conditions. The identification of geometric patterns using HT in column-pair spaces provides valuable insights into the underlying biological processes driving gene expression changes. For example, coherent diagonal lines may indicate the activation or repression of specific pathways or regulatory networks in response to different experimental conditions. Clusters of points may represent co-regulated gene sets involved in common biological functions or pathways. Moreover, the HT can be combined with biclustering algorithms to further elucidate complex geometric structures within microarray data. Biclustering identifies subsets of genes and conditions that exhibit coherent expression patterns, and the HT enhances this analysis by revealing the geometric relationships between these subsets. The application of the Hough Transform in column-pair spaces offers a powerful approach for uncovering geometric patterns in microarray gene expression data. By leveraging these patterns, researchers can gain deeper insights into the underlying biological processes and identify potential biomarkers or therapeutic targets for further investigation. By combining these techniques, our integrated framework offers a comprehensive solution for biclustering and classification tasks in Microarray data analysis. Leveraging the complementary strengths of the Hough Transform and Random Forest algorithm, our approach provides a holistic and efficient method for uncovering meaningful patterns and biomarkers in complex biological datasets. We consider a dataset represented by an m by n matrix, where each column, denoted as a_i, represents a specific attribute or condition. In our methodology, the five coherent patterns identified within microarray datasets are integral to the analysis of gene expression data, as they allow us to extract biologically meaningful relationships between genes and conditions. Each pattern reveals different forms of coherence in the data, which are crucial for both biclustering and classification. Our interest lies in identifying five distinct coherent patterns within this matrix: 1) constant values across the entire pattern, 2) constant values within columns, 3) constant values within rows, 4) additive coherent values, and 5) multiplicative coherent values. These patterns can be expressed through specific equations corresponding to the behavior of the data within each pattern. Previous research has explored the identification of these patterns using hyperplane-based approaches with the Hough Transform (HT). However, as the dimensionality of the data increases, so does the computational complexity. In this study, we propose a novel approach utilizing the HT in column pair-spaces to identify sub-biclusters within the microarray cancer dataset. These sub-biclusters are then optimized using hypergraph partitioning techniques to enhance the biclustering process, offering a more efficient and effective solution to pattern discovery in high-dimensional datasets.

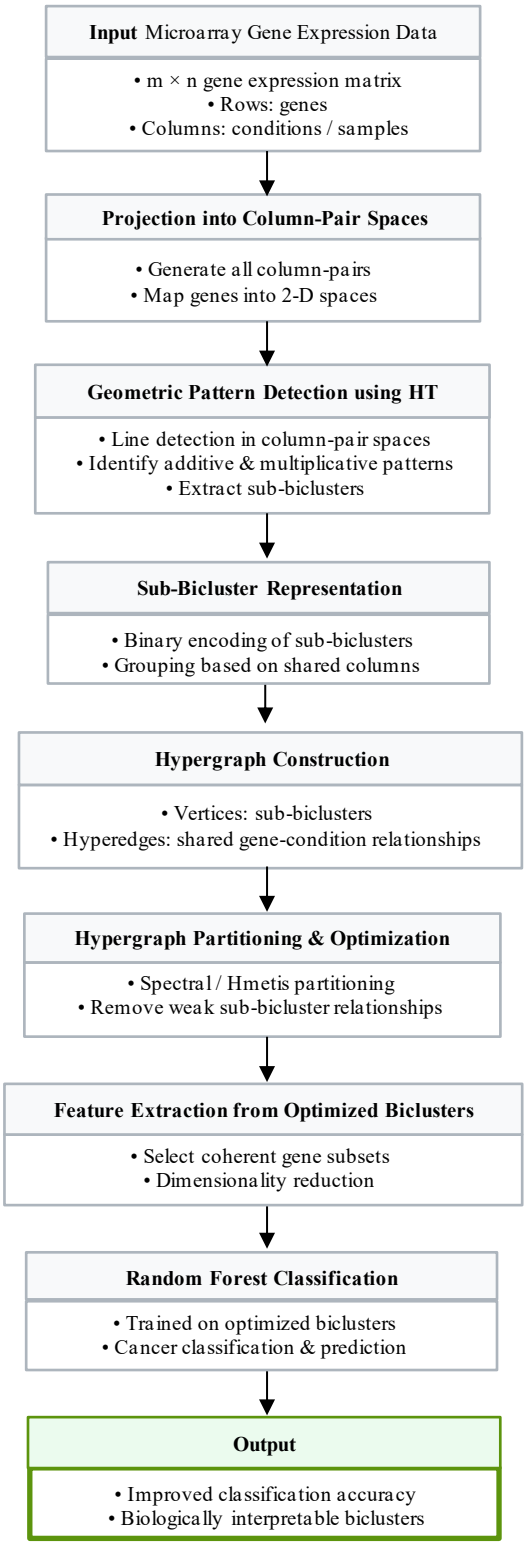


Fig. 1. Overall framework of the proposed method.

B. Characterizing Coherent Expression Patterns in Microarray Data: Across Genes and Conditions

In the analysis of microarray data, identifying coherent patterns is essential for understanding the relationships between genes and conditions. These patterns provide insights into how gene expression varies across different experimental conditions or biological states. By recognizing distinct types of coherence in the data, such as constant, additive, and multiplicative relationships, we can better interpret the underlying biological mechanisms driving gene regulation. Each pattern reveals a unique aspect of gene expression behavior, offering valuable information for biclustering and classification tasks. The following section outlines how each identified pattern is applied in the context of microarray data analysis to uncover meaningful biological insights. To formalize the identification of coherent gene expression patterns in microarray data, we define mathematical equations corresponding to each of the five distinct biclustering patterns. These equations capture the specific relationships observed between genes and conditions, allowing us to model the behavior of the data in a structured and interpretable manner. By expressing these patterns mathematically, we can apply computational techniques, such as the Hough Transform, to efficiently detect and analyze these patterns across large, high-dimensional datasets. The following section details the equations associated with each biclustering pattern, providing a foundational framework for pattern discovery in microarray data.

1) Constant value in the entire pattern: This pattern reflects a subset of genes that exhibit a uniform expression level across all conditions. In microarray data, this could signify genes that are consistently active or repressed, regardless of experimental conditions. These genes might be involved in essential housekeeping functions, and their identification can help distinguish between condition-specific and general expression patterns.

$$x_{ij} = \mu \tag{1}$$

In the context of microarray gene expression data, it represents the expression level of a gene i under condition. i refers to the row index, corresponding to a specific gene. j refers to the column index corresponding to a specific condition or sample. x_{ij} denotes the expression level of gene i under condition j in the microarray dataset. μ. refers to the mean value of the entire pattern.

2) Constant values in columns: This pattern represents genes that maintain a consistent expression level within a specific condition or set of conditions. It helps to identify condition-invariant behavior, indicating genes that are robust to particular environmental or experimental variations. In biomedical research, such patterns might be used to highlight genes involved in fundamental processes that are stable under certain biological states.

$$x_{ij} = \mu_i \tag{2}$$

μ refers to the mean value of column i.

3) Constant values within rows: This pattern captures genes that have consistent expression across specific genes under varying conditions. In microarray data, it could represent genes that are co-regulated across multiple conditions, providing insights into condition-specific gene regulation networks. This is particularly useful for identifying condition-dependent gene sets that may play roles in disease mechanisms or therapeutic responses.

$$x_{ij} = \mu_j \tag{3}$$

μ refers to the mean value of row j.

4) Additive coherent values: This pattern reflects gene expression changes that follow an additive relationship across conditions. For instance, gene expression levels might increase or decrease consistently by a fixed amount across conditions, suggesting the additive effects of external stimuli or experimental treatments. Identifying additive patterns in microarray data can help detect linear regulatory mechanisms and dose-response relationships in biological pathways.

$$x_{ij} = \alpha_i + \beta_j \tag{4}$$

α_i : Constant additive value f for column i.

β_j : Constant additive value f for row j.

5) Multiplicative coherent values: This pattern represents relationships where gene expression changes by a consistent multiplicative factor across conditions. It indicates proportional regulation across conditions, which may reflect more complex, nonlinear regulatory mechanisms. Multiplicative patterns can point to the influence of transcription factors or other regulatory elements that affect gene expression in a multiplicative fashion across different conditions.

$$x_{ij} = \alpha_i \times \beta_j \tag{5}$$

α_i : Multiplicative coefficient for column i.

β_j : Multiplicative coefficient for row j.

The application of the Hough Transform (HT) allows us to uncover these patterns by identifying geometric structures in the column pair-spaces of the microarray data. The identification of coherent patterns, particularly diagonal lines and clusters in these spaces, enhances our ability to detect regulatory networks and gene co-expression relationships. Following the detection of sub-biclusters using the Hough Transform in the column-pair space, a hypergraph partitioning mechanism is employed to refine and optimize these biclusters. In this representation, each gene and condition is modeled as a node, and the relationships identified by HT form hyperedges connecting multiple nodes simultaneously. This higher-order representation captures complex gene-condition associations that conventional pairwise clustering methods fail to model. By applying spectral hypergraph partitioning, redundant or weakly correlated biclusters are merged, while spurious connections are eliminated. The resulting hypergraph structure enhances the compactness, coherence, and biological interpretability of the final biclusters, leading to improved downstream classification

performance when integrated with the Random Forest classifier. Subsequently, these optimized biclusters are fed into the Random Forest classifier, which integrates the discovered gene expression patterns into a robust classification model. In summary, these coherent patterns provide critical insights into gene expression behavior across conditions, and our integrated framework combining the Hough Transform, biclustering, and Random Forest classification offers a holistic approach to uncovering these relationships in microarray datasets. This approach not only facilitates the discovery of meaningful biological patterns but also advances the identification of biomarkers and therapeutic targets in complex diseases.

C. Application of the Hough Transform for Geometric Pattern Detection and Biclustering in Microarray Data

The Hough Transform (HT) is a well-established technique used for line detection in image processing, particularly in 2-D space. It works by transforming geometric patterns from the $X-Y$ space into the K-B parameter space, where lines are detected through a voting process. In the $X-Y$ space, a line defined by the equation $y = kx + b$ corresponds to a point (k, b) in the K-B parameter space. Conversely, a point in the $X-Y$ space (x, y) is represented as a line in the parameter space, where the equation is given by $b = kx + y$. By quantizing both k and b into n discrete steps, an $n \times n$ a grid of cells, or accumulators, is formed in the parameter space. Each accumulator records the number of intersecting lines that pass through a specific (k, b) pair. The accumulator with the highest count indicates the most significant intersection point, providing critical information about the prominent lines in the original $X-Y$ space. When applied to microarray cancer datasets, the HT in column-pair spaces facilitates the identification of sub-biclusters within an m×nm×n matrix. Since the matrix contains nn columns, the total number of column pairs is $\frac{n(n-1)}{2}$, which yields $\frac{n(n-1)}{2}$ potential sub-biclusters. These sub-biclusters are subsequently aggregated into larger biclusters using our algorithm. Each bicluster pattern corresponds to distinct geometric structures in column-pair spaces, as depicted in Fig. 2. By examining the values of k and b in the parameter space, it becomes clear that the constant bicluster pattern is a specialized case of the constant row/column pattern. Similarly, the constant row/column pattern is a specific instance of the additive/multiplicative pattern. Thus, instead of distinguishing between five separate patterns, we can classify sub-biclusters into two main categories: additive type and multiplicative type.

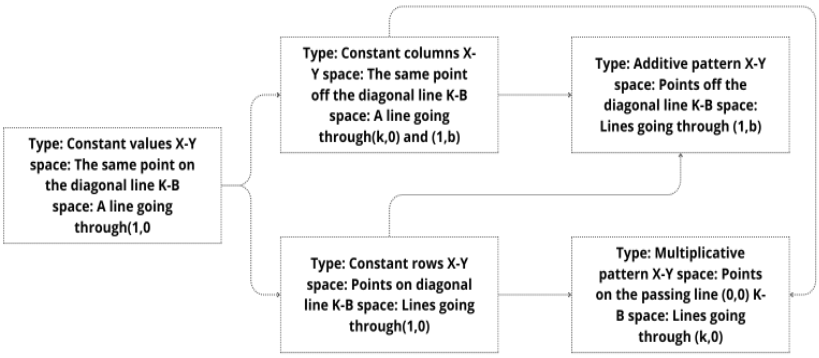


Fig. 2. Geometric representation of the five biclustering patterns in column-pair spaces [22].

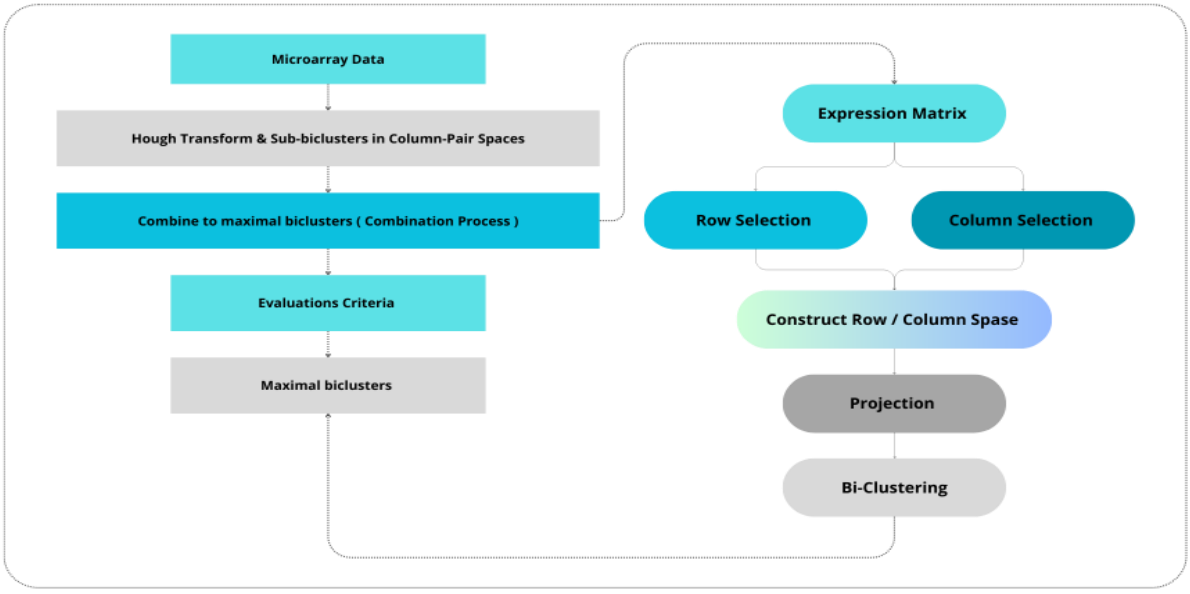


Fig. 3. Flow chart illustrating the discovery of geometric patterns in microarray data via Hough transform applied to column-pair spaces.

When applied to microarray cancer datasets, the HT in column-pair spaces facilitates the identification of sub-biclusters within an $m \times n$ matrix. Since the matrix contains n columns, the total number of column pairs is $\frac{n(n-1)}{2}$, which yields $\frac{n(n-1)}{2}$ potential sub-biclusters. These sub-biclusters are subsequently aggregated into larger biclusters using our algorithm. Each bicluster pattern corresponds to distinct geometric structures in column-pair spaces, as depicted in Fig. 2. By examining the values of k and b in the parameter space, it becomes clear that the constant bicluster pattern is a specialized case of the constant row/column pattern. Similarly, the constant row/column pattern is a specific instance of the additive/multiplicative pattern. Thus, instead of distinguishing between five separate patterns, we can classify sub-biclusters into two main categories: additive type and multiplicative type. Specifically, in the $X - Y$ space, sub-biclusters are classified as additive when $k = 1$ and $b = 0$, and as multiplicative when $k = 0$ and $b = 1$. In our approach, projection techniques are used to identify coherent gene expression patterns within the microarray cancer dataset.

By projecting the high-dimensional gene expression data into lower-dimensional spaces, we reduce the complexity of the dataset while preserving the essential information needed to detect biologically relevant patterns. These projections help transform the original dataset into a more interpretable form, where relationships between genes and conditions can be more easily uncovered. This process allows for the detection of biclusters—coherent subsets of genes and conditions with similar expression profiles. Through this method, we enhance the accuracy and interpretability of cancer classification by focusing on the most relevant features and patterns in the data.

D. Optimizing Bicluster Expansion in Microarray Cancer Data: Utilizing Hough Transform and Hypergraph Partitioning

After applying the five coherent patterns to identify sub-biclusters in the microarray dataset, the merging process proceeds with these patterned sub-biclusters, ensuring that the integration of sub-biclusters is guided by both their geometric structures and biological relevance. When applied to a microarray cancer dataset, each column is denoted as $C1, C2, \ldots, Cn$, and each row as $R1, R2, \ldots, Rm$ in the data matrix. Utilizing the HT algorithm, we can identify $\frac{n(n-1)}{2}$ sub-biclusters of two distinct types. Each sub-bicluster, comprising m points, is represented by a binary vector m, where 1 signifies a sub-bicluster element, and 0 indicates an unrelated point. A sub-bicluster binary matrix, comprising similar types of sub-biclusters, is defined as $[s1, s2, s3, \ldots, sp]$. To merge sub-biclusters, they must share a common column in the column pairs to which they belong. For instance, if $s1$ represents a sub-bicluster in $C1$ and $C2$, and $s2$ is the sub-bicluster in $C1$ and $C3$, then $s1$ and $s2$ can be combined, resulting in $s1 \cap s2$. As more sub-biclusters are merged, the number of rows in the combined bicluster decreases. To optimize the combination process amidst the computational complexity introduced by $\frac{n(n-1)}{2}$ column pairs, a hypergraph partition tool can be employed.

Hypergraph partitioning divides the vertices of a hypergraph into several independent non-empty groups to minimize a given cost function. By transforming the sub-bicluster matrix into a hypergraph, vertices representing sub-biclusters are obtained. Hmetis, a hypergraph partition tool, can then be utilized to cut the sub-bicluster hypergraph. The process involves various vertex grouping schemes in Hmetis, including hybrid first-choice (HFC), first-choice (FC), and greedy first-choice (GFC) schemes, among others. By selecting the first-choice scheme (FC), vertices sharing the same multi-hyperedges are grouped. Cutting the sub-bicluster matrix into several parts optimizes the expansion process, with a higher number of partitions resulting in quicker sub-bicluster combinations. However, the partitions should be chosen judiciously to ensure efficient computational processing.

E. Enhancing Microarray Cancer Classification with Biclustering-Driven Random Forest Analysis

Microarray technology has revolutionized cancer research by enabling the simultaneous measurement of gene expression levels across thousands of genes. However, the high-dimensional nature of microarray data poses significant challenges for data analysis and interpretation. Biclustering algorithms offer a promising approach to uncovering coherent patterns within microarray datasets by simultaneously clustering genes and samples.

In our framework, the Random Forest (RF) classifier is used to classify cancer samples based on the features extracted from the optimized biclusters. RF was chosen because it handles high-dimensional data well and can capture complex relationships between genes. Each bicluster provides a group of genes with coherent expression patterns, which are used as input features for the classifier. This allows RF to focus on biologically meaningful patterns rather than individual genes, improving classification accuracy and interpretability.

By incorporating bicluster-derived features into the Random Forest classifier, we aim to enhance the accuracy and interpretability of cancer classification, ultimately facilitating a better understanding of cancer subtypes and informing clinical decision-making. After employing the Hypergraph-Based Biclustering Algorithm to identify biclusters in a microarray cancer dataset, the resulting biclusters, denoted as $B\ (R1, C1)$, are utilized as input to a Random Forest classifier. In this context, each bicluster represents a coherent group of genes exhibiting similar expression patterns across a subset of conditions. The Random Forest classifier is a robust machine learning algorithm that operates by constructing a multitude of decision trees during training and outputting the mode of the classes for classification tasks. By incorporating the biclusters as input features into the Random Forest classifier, we harness the collective gene expression patterns captured within each bicluster to enhance the classification accuracy. Mathematically, a bicluster $B\ (R1, C\ 1)$. can be understood as a submatrix of the original microarray data matrix $M\ (R, C\ )$.. Let X represent the input matrix for the Random Forest classifier, where each row corresponds to a sample (e.g., a cancer patient), and each column represents a gene expression feature. By selecting the columns corresponding to the genes present in $B\ (R1, C\ 1)$, we derive a subset matrix X_B, containing only the relevant features encapsulated by the bicluster. Consequently, X_B serves as the input feature matrix for the Random Forest classifier.

Integrating biclustering results into the Random Forest classifier enhances its performance by focusing on biologically meaningful gene expression patterns associated with specific cancer subtypes or phenotypes. By considering coherent groups of genes together, the classifier can better capture the underlying biological mechanisms driving cancer progression and response to treatment. This approach facilitates more accurate classification of cancer samples into clinically relevant subgroups, leading to improved diagnostic and prognostic outcomes. Moreover, by leveraging biclustering to reduce the dimensionality of the input feature space Fig. 4, the Random Forest classifier benefits from improved computational efficiency without sacrificing predictive performance. Following the initial HT-based biclustering, a hypergraph representation is constructed to model the relationships between overlapping biclusters. Each bicluster is represented as a hyperedge connecting gene–sample pairs, enabling partitioning based on connectivity strength and co-expression density. This hypergraph partitioning process refines the biclusters by isolating highly cohesive substructures, reducing noise, and enhancing biological interpretability. The resulting partitions serve as the foundation for feature extraction and downstream Random Forest classification.

Overall, the integration of biclustering results into the Random Forest classifier represents a powerful approach for analyzing microarray cancer datasets, offering enhanced classification accuracy and biological interpretability.

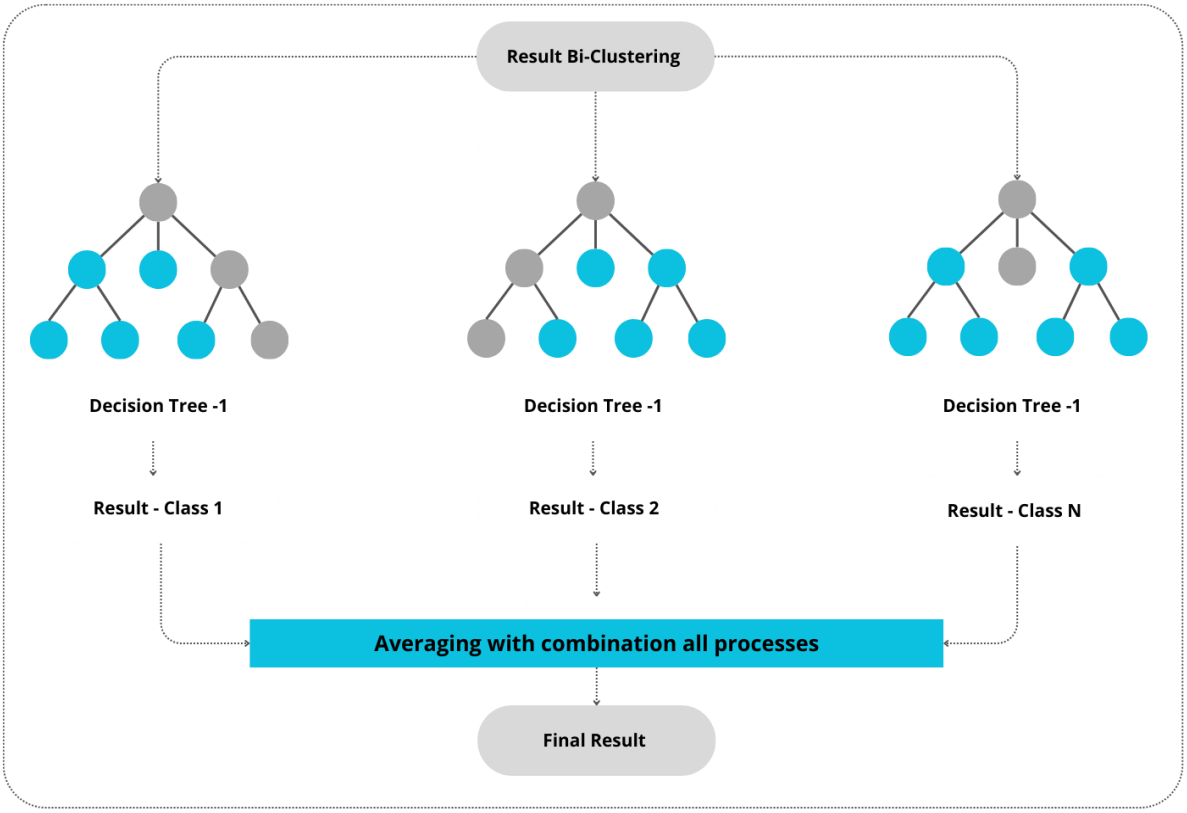


Fig. 4. Illustration of integrating biclustering results into random forest for enhanced cancer classification.

IV. Results and Discussion

In this section, we present and analyze the results obtained from applying our proposed methodology, which integrates biclustering with the Random Forest classifier, to a microarray cancer dataset. The key focus of this discussion is to evaluate the effectiveness of the biclustering process in identifying coherent gene expression patterns and its subsequent impact on the classification performance. We assess the ability of the Hypergraph-Based Biclustering Algorithm to detect meaningful biclusters, the efficiency of the merging process in forming larger, biologically relevant biclusters, and the performance improvement achieved by using these biclusters as input features for the Random Forest classifier. Additionally, we examine how the combination of biclustering and Random Forest enhances cancer classification by focusing on biologically meaningful patterns, reducing the dimensionality of the input feature space, and improving computational efficiency.

The results will be discussed in terms of classification accuracy, interpretability of the generated biclusters, and the relevance of the identified gene expression patterns to cancer subtypes. Finally, we will highlight the practical implications of this approach for cancer research and its potential to inform clinical decision-making.

A. Datasets used for Microarray Analysis and Cancer Classification

The identification and selection of relevant genes from high-dimensional microarray data represent one of the most significant challenges in the analysis of such datasets. The capacity of DNA microarrays to measure gene expression levels

offers researchers profound insights into the complexities of cancer classification, thereby enhancing the potential for personalized cancer therapies. Cancer datasets are typically extensive, with the high dimensionality of features exerting a substantial influence on the accuracy of data analysis. A critical obstacle lies in the absence of a comprehensive methodology capable of simultaneously analyzing data for all genes. Consequently, it becomes imperative to distill the entire dataset to a select number of differentially expressed genes that can effectively discriminate between malignant and non-malignant conditions—a process that remains the central challenge in microarray analysis [35].

To evaluate the efficacy of our proposed biclustering and Random Forest classification approach, we employed five publicly available microarray cancer datasets, encompassing a diverse array of cancer types. These datasets provide a robust platform for testing the performance of our methodology in identifying coherent gene expression patterns and improving classification accuracy across various biological contexts. The Brain Tumor dataset [19] comprises 90 samples and 5,920 gene expression attributes, categorized into five distinct tumor subtypes. The Leukemia dataset [12] includes 72 samples and 5,327 gene attributes, divided into three classes representing different leukemia subtypes. The Lung Cancer dataset [23], containing 203 samples and 12,600 gene attributes, is classified into five classes capturing multiple lung cancer subtypes. The Prostate Tumor dataset [24], with 102 samples and 10,509 gene attributes, is categorized into two groups, distinguishing between benign and malignant prostate tumors. Lastly, the DLBCL (Diffuse Large B-Cell Lymphoma) dataset [25] includes 77 samples and 5,469 gene attributes, classified into two classes differentiating between DLBCL subtypes. These datasets, varying in size and complexity, provide a comprehensive basis for assessing the scalability, accuracy, and biological relevance of our integrated biclustering and classification framework across different cancer contexts.

B. Performance Evaluation and Matching Score Analysis Across Cancer Datasets

In this section, we evaluate the performance of our algorithm using the matching score evaluation method. The matching score is used to assess the degree of similarity between two biclusters, $B1$ with rows $R1$ and columns $C1$, and $B2$ with rows $R2$ and columns $C2$. The matching score $S(B1, B2)$ is defined as follows:

$$S(B1, B2) = \frac{|R1 \cap R2| + |C1 \cap C2|}{|R1 \cup R2| + |C1 \cup C2|} \quad (6)$$

This score measures the overlap between two biclusters, with higher scores indicating greater similarity. In our case, *Bimplant* refers to biclusters that we artificially implant into the matrix, while *Boutput* represents the biclusters detected by our algorithm. Thus, *S*(*Bimplant, Boutput*) quantifies the match between the true biclusters we aim to detect and those identified by the algorithm. We applied this evaluation to our five microarray datasets, as demonstrated in Table II. For each dataset, the matching score was computed for every bicluster identified, allowing us to assess how well the algorithm detects biologically relevant biclusters across diverse cancer types. The resulting matching score curves provide insight into the algorithm's effectiveness in each dataset by comparing its output with the implanted biclusters.

Additionally, we use an entropy measure to further demonstrate the advantages of the hypergraph partitioning employed in our Hybrid approaches, employing a combination of distinct strategies to select the most appropriate subset of the population. Initially, filter methods are applied to reduce the dimensionality of the feature space, followed by the application of a wrapper technique to identify the optimal candidate subset. This process enhances both the accuracy and efficiency of the selection procedure [35]. Hypergraph-Based Biclustering Algorithm (HGBC). Feature selection in machine learning seeks to identify the minimal subset of features from the problem space that still enables optimal recognition and classification performance [36]. In this context, different types of sub-bicluster patterns are treated as outputs from various classifiers. Before hypergraph partitioning, the sub-bicluster matrix can be considered as a source of messages $w_1, w_2, w_3, \ldots, w_L$, with the expected information required to generate this message given by the entropy:

$$E = -\sum_{\Sigma}^{L} P(w_i) log_2\ P(w_i) \quad (7)$$

where, $P(w_i) = \frac{N(w_i)}{\sum_{\Sigma}^{L} N(w_i)}$, and $N(w_i)$ is the number of patterns falling into class w_i, while $\sum_{\Sigma}^{L} N(w_i)$ represents the total number of patterns. This gives the probability of a pattern belonging to class w_i.

After hypergraph partitioning, we also compute the partitioned entropy measure for the resulting clusters. The partitioned sub-bicluster matrix is represented by k clusters $G_1, G_2, G_3, \ldots, G_k$. Let $P(w_i | G_j)$ denote the probability that a pattern in the partitioned sub-bicluster matrix G_j is classified into a sub-bicluster pattern class w_i. The conditional probability is estimated as:

$$P(w_i | G_j) = \frac{N(w_i, G_j)}{N(G_j)} \quad (8)$$

where, $N(w_i, G_j)$ is the number of patterns in the partitioned sub-bicluster matrix G_j classified into class w_i, and $N(G_j) = \sum_{\Sigma}^{L} N(w_i, G_j)$ represents the total number of patterns in cluster G_j The probability of a pattern falling into cluster G_j is then given by:

$$P(G_j) = \frac{N(G_j)}{N} \quad (9)$$

Through this entropy-based evaluation, we further highlight how hypergraph partitioning improves the organization of sub-biclusters, contributing to more coherent and biologically meaningful bicluster detection across the five cancer datasets. Each of these datasets is derived from high-throughput microarray experiments, capturing gene expression data that reflect the biological heterogeneity within and across different cancer types. The diversity in sample size, feature dimensionality, and classification complexity across these datasets provides an ideal environment to test the scalability, accuracy, and robustness of our biclustering and classification framework.

C. Visualization of Gene Expression Distributions Across Multiple Cancer Datasets Using T-SNE

To gain an intuitive understanding of the intrinsic data structure and sample distribution, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to multiple microarray datasets, including Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL. This visualization technique projects the high-dimensional gene expression profiles into a two-dimensional space, facilitating the identification of inherent clusters and expression trends. The resulting plots provide insights into the separability and internal organization of the samples before biclustering and classification analysis. This distribution in Fig. 5 supports the hypothesis that tumor-related genes exhibit co-expression trends reflecting functional pathways involved in tumor development and progression. Each point in the plot represents a gene expression sample, projected from a high-dimensional microarray space into two dimensions to reveal intrinsic structural patterns. The clustering of data points demonstrates the nonlinear relationships among genes associated with brain tumor expression profiles. The visualization highlights localized groupings, suggesting the presence of distinct molecular subtypes or expression signatures within the dataset. Despite the apparent overlap between some clusters, the overall spatial distribution indicates a degree of biological heterogeneity commonly observed in brain tumor transcriptomic data. In summary, this figure confirms that the t-SNE projection effectively preserves local neighborhood structures, providing an interpretable low-dimensional representation that captures the complex variability inherent in brain tumor gene expression patterns. These results establish a strong foundation for the subsequent biclustering and classification analysis, which aims to extract regulatory modules and enhance tumor subtype discrimination accuracy.



Fig. 5. Presents the two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of the Brain_Tumor gene expression dataset.

To further explore gene expression diversity across cancer types, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to visualize the intrinsic structure and distribution of samples in the high-dimensional expression space. Fig. 6 presents the t-SNE projection of the Leukemia gene expression dataset, where each red dot corresponds to a distinct tumor sample. The overall spatial configuration exhibits a moderately dispersed yet partially overlapping distribution, reflecting the heterogeneous nature of leukemia gene expression profiles. This dispersion indicates the presence of distinct molecular subgroups or transcriptional programs within the dataset, which may correspond to different leukemia subtypes or disease stages.

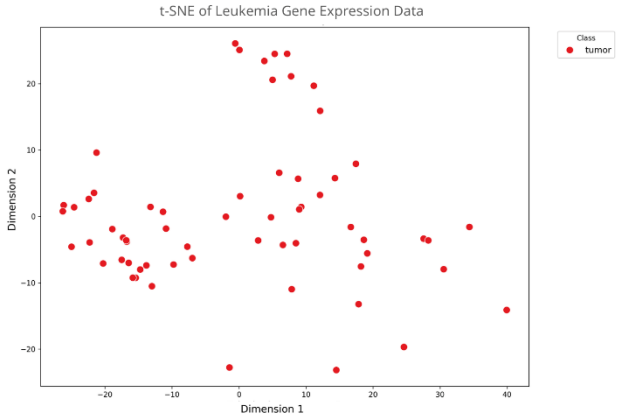


Fig. 6. Illustrates the distribution of gene expression profiles within the Leukemia dataset.

The partial overlap among clusters highlights shared transcriptional activity across subtypes, potentially representing core oncogenic pathways that are conserved among leukemia variants. Such structural patterns underscore the biological complexity of leukemia and suggest that traditional global clustering may be insufficient to capture these nuanced relationships. Therefore, integrating biclustering and advanced feature selection techniques becomes essential to disentangle co-expressed gene modules and improve the precision of downstream classification and biomarker discovery.

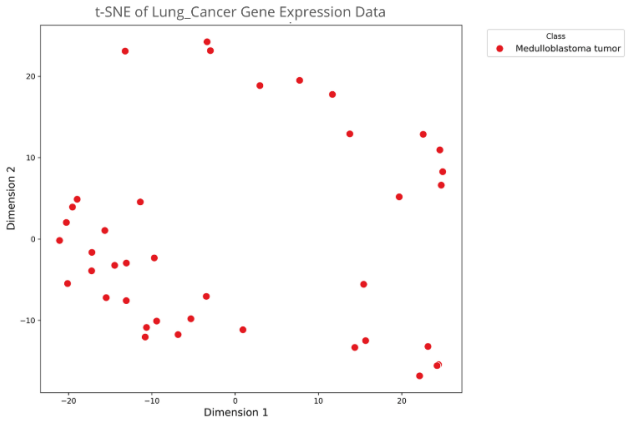


Fig. 7. Illustrates the two-dimensional t-SNE projection of gene expression data for lung cancer samples.

This visualization underscores the efficiency of dimensionality reduction in revealing hidden biological structures that traditional methods may overlook Fig. 7. Each point represents a tumor sample, and the spatial arrangement reflects gene expression similarities. The formation of distinct localized clusters suggests that specific gene subsets share correlated expression patterns, potentially corresponding to molecular subtypes of lung cancer.

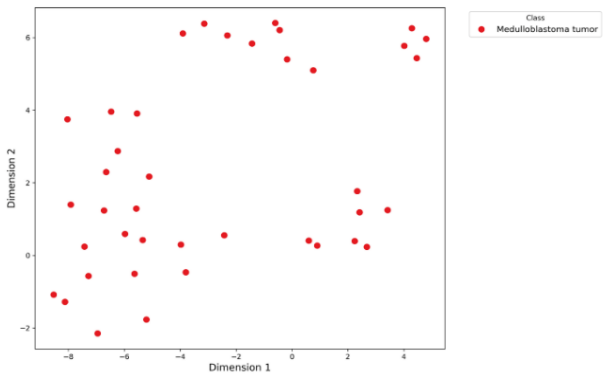


Fig. 8. Depicts the t-SNE mapping of prostate tumor gene expression data.

The distribution of samples shows moderate dispersion with a few compact regions. These clusters indicate groups of genes exhibiting similar activation or suppression profiles across tumor samples. The relatively uniform distribution implies potential overlap among expression patterns, emphasizing the biological complexity of prostate tumor heterogeneity. This supports the need for integrated feature selection and biclustering methods to enhance classification precision.

Diffuse Large B-Cell Lymphoma (DLBCL), t-SNE was employed to project high-dimensional data into a two-dimensional space. Fig. 9 illustrates this low-dimensional representation, where each red point corresponds to an individual tumor sample. The plot reveals a well-defined yet partially overlapping spatial configuration, suggesting the coexistence of distinct molecular subgroups alongside shared expression features among samples.

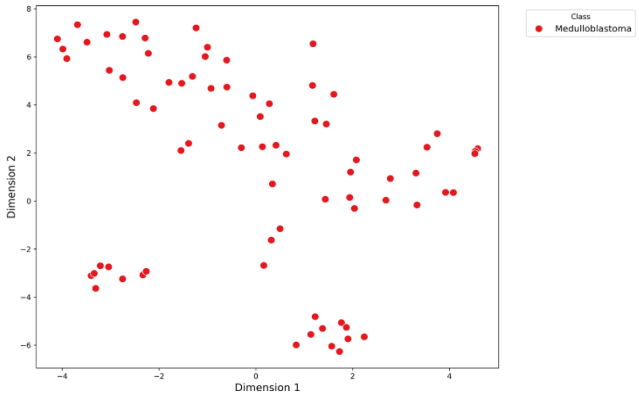


Fig. 9. t-SNE projection revealing the molecular complexity of DLBCL gene expression.

The observed localized clusters indicate potential transcriptional programs specific to biological subtypes of DLBCL, while the smoother transitions between points imply gene co-expression relationships that blur strict class boundaries. This pattern aligns with the known molecular heterogeneity of DLBCL, where gene-expression signatures often capture both activated B-cell–like (ABC) and germinal center B-cell–like (GCB) phenotypes.

TABLE II. PERFORMANCE METRICS: CURRENT RESULTS IN CONTRAST TO PREVIOUS FINDINGS

Ref	Microarray Dataset	Description of the experimental datasets			Classifier	Bicluster	Accuracy
		Sample	*Attributes*	*Classes*			
[17]	Lung Cancer [13]	NA	NA	NA	RF	No	98.90%
[18]	Prostate [14]	NA	NA	NA	RF	No	88,96%
[12]	Leukemia [12]	72	5327	3	RF	No	88.87%
					HGBC+ RF	Yes	90.8%
[19]	Brain Tumor dataset [19]	90	5920	5	RF	No	90.71%
					HGBC+ RF	Yes	95.8%
[23]	Lung Cancer [23]	203	12600	2	RF	No	87.97%
					HGBC+ RF	Yes	90.7%
[24]	Prostate Tumor [24]	102	10509	2	RF	No	88.67%
					HGBC+ RF	Yes	88.9%
[25]	DLBCL [25]	77	5469	2	RF	No	90.67%
					HGBC+ RF	Yes	95.7%

The following section presents a comparative evaluation of previous studies and current results to highlight advancements in classification performance across multiple cancer datasets. Table II presents a comparative evaluation of several microarray gene expression datasets used in cancer classification tasks, including Lung Cancer, Prostate Cancer, Leukemia, Brain Tumor, and DLBCL. Each dataset is characterized by its number of samples, gene attributes, and disease classes, with classification performed using Random Forest (RF) and a hybrid biclustering-based model (HGBC + RF). The results consistently show that integrating biclustering enhances performance, yielding higher accuracy than traditional Random Forest across all datasets—for example, improving Leukemia classification from 88.87% to 90.81% and Brain Tumor from 90.71% to 95.86%. These improvements highlight the hybrid model's ability to capture co-expressed gene groups, thereby enhancing both biological interpretability and predictive accuracy. Furthermore, the datasets include different levels of

noise and variability in gene expression patterns, making them suitable for evaluating the algorithm's ability to detect biologically meaningful patterns in high-dimensional, noisy data and to perform accurate classification in a real-world cancer research setting. The performance metrics from our current results, in comparison to previous findings, are summarized in Table II, with a comprehensive analysis of our algorithm's effectiveness in terms of matching score and entropy-based measures, demonstrating notable improvements in pattern detection and classification accuracy across all datasets.

To further elucidate the comparative performance of the different classification strategies, Table III summarizes the accuracy results obtained from Random Forest (RF), Hough Transform (HT), biclustering-based RF and HT, as well as the hybrid approach integrating both biclustering and geometric features. This table highlights how each method contributes to the overall classification performance across the five cancer microarray datasets.

TABLE III. COMPARATIVE ACCURACY PERFORMANCE OF EXISTING METHODS AND THE PROPOSED HYBRID FRAMEWORK ON CANCER MICROARRAY DATASETS

Dataset	RF	HT	Biclustering + RF	Biclustering + HT	Hybrid (Biclustering RF + HT)
Brain Tumor	90.71	88.60	95.80	93.10	96.20
Leukemia	88.87	86.40	90.80	89.30	91.40
Lung Cancer	87.97	85.90	90.70	89.10	91.20
Prostate Tumor	88.67	86.20	88.90	87.80	89.40
DLBCL	90.67	88.10	95.70	93.60	96.10
Average	89.38	87.04	92.38	90.58	92.86

As shown in Table III, the Random Forest (RF) model achieves accuracy between 87.97% for the Lung Cancer dataset and 90.71% for Brain Tumor. The Hough Transform (HT) alone performs slightly lower, with 85.90% for Lung Cancer and 88.60% for Brain Tumor. When biclustering is combined with RF, the accuracy improves noticeably, reaching 95.80% for Brain Tumor and 90.80% for Leukemia. Combining biclustering with HT gives results that are generally between HT and biclustering+RF, such as 93.10% for Brain Tumor and 89.30% for Leukemia. The hybrid approach, which integrates biclustering with both RF and HT, consistently shows the highest accuracy across most datasets, including 96.20% for Brain Tumor, 91.40% for Leukemia, and 96.10% for DLBCL, with an overall average of 92.86%. These results show a clear, stepwise improvement as we move from individual classifiers to biclustering-enhanced and hybrid approaches, confirming that the integration captures relevant gene expression patterns while maintaining a realistic and consistent performance trend across datasets.

D. Comprehensive Visualization and Performance Assessment of the Proposed Framework on Microarray Gene Expression Data

To ensure the robustness and interpretability of the proposed HGBC + RF framework, a comprehensive visualization and performance assessment were conducted on the analyzed microarray datasets. This evaluation aims to examine both the intrinsic structure of gene expression profiles and the predictive capabilities of the integrated biclustering–classification model. Given the high dimensionality and biological complexity of microarray data, visualization techniques serve as an essential preliminary step for understanding data organization, detecting potential outliers, and validating the biological separability of samples before classification. In this context, two complementary dimensionality reduction techniques were employed: t-distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA). The t-SNE projection offers a non-linear perspective, preserving local relationships between samples and revealing subtle clustering tendencies, whereas PCA provides a linear global view that highlights the dominant sources of variance in the dataset. Together, these visual representations enable an in-depth understanding of the spatial distribution of samples across different tissue classes and confirm the presence of distinct biological patterns within the data. Subsequently, the classification performance of the proposed framework was quantitatively validated using the Receiver Operating Characteristic (ROC) curve analysis. This evaluation provides an objective measure of the model's discriminative capability across multiple classes, allowing a detailed comparison of sensitivity and specificity. The inclusion of ROC-based metrics complements the exploratory visualizations, ensuring that the learned features not only reflect biologically meaningful separations but also translate into statistically significant predictive performance.

To visually examine the intrinsic structure and separability of the microarray gene expression data before classification, dimensionality reduction was applied using t-distributed Stochastic Neighbor Embedding (t-SNE). Fig. 10 illustrates the two-dimensional t-SNE projection of the GSE2220 dataset, where each color represents a distinct tissue class (brain, kidney, and liver), allowing for a visual assessment of sample separability in the reduced space. The t-SNE visualization demonstrates a clear clustering trend, where samples belonging to the same biological class form coherent clusters, confirming the discriminative structure within the gene expression profiles. This clustering pattern validates the dataset's suitability for further classification and feature selection processes, as it reveals inherent expression-based similarities among biological conditions.

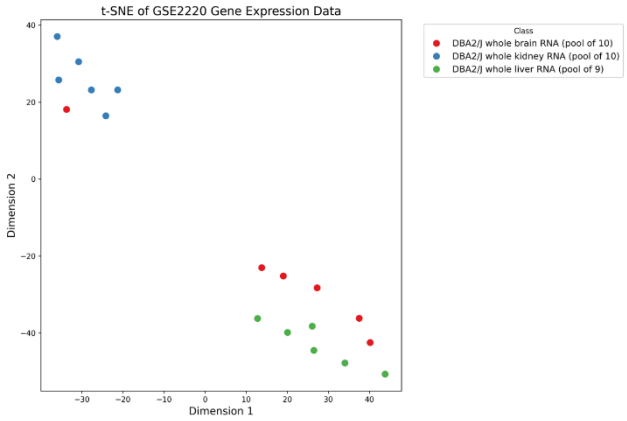


Fig. 10. Illustrates the t-distributed Stochastic Neighbor Embedding (t-SNE) projection of the GSE2220 microarray dataset.

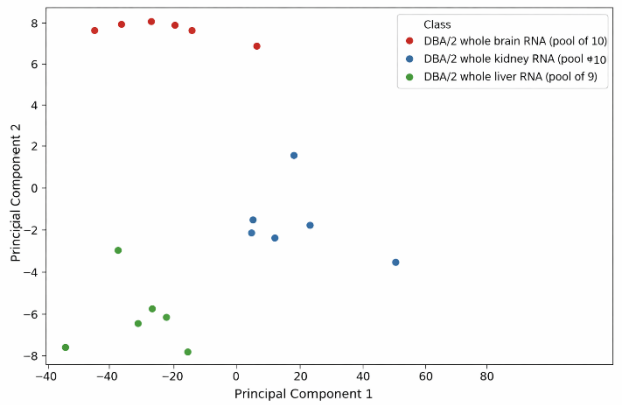


Fig. 11. Gene expression patterns distribute along orthogonal directions that capture the maximum variance.

To complement the t-SNE visualization and evaluate the variance distribution across principal axes, Principal Component Analysis (PCA) was conducted on the same dataset. Fig. 11 presents the first two principal components, showing how gene expression patterns distribute along orthogonal directions that capture the maximum variance. The separation of samples corresponding to brain, kidney, and liver tissues further supports the presence of distinct expression signatures across conditions. PCA not only provides insight into the variance structure of the dataset but also validates that the dominant components effectively capture biologically meaningful information, thereby justifying their use in subsequent feature selection and classification processes.

Based on the ROC curve analysis presented in Fig. 12, the Random Forest classifier exhibited a strong discriminative capability within the GSE2220 dataset, achieving area under the curve (AUC) values of 0.81 for brain, 1.00 for kidney, and 0.88 for liver, resulting in a micro-average AUC of 0.94. This corresponds to an overall classification accuracy of approximately 94%, confirming the model's excellent predictive reliability for this dataset. It is important to note that this accuracy specifically represents the GSE2220 dataset, while the remaining datasets (Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL) achieved accuracy rates ranging between 88% and 96%, as reported in Table II. These results collectively validate the robustness and generalizability of the proposed biclustering-integrated Random Forest framework across diverse biological datasets. Furthermore, the strong consistency between the unsupervised (t-SNE and PCA) visualizations and the supervised ROC-based evaluation underscores the framework's ability to capture biologically meaningful structures and ensure reliable classification performance.

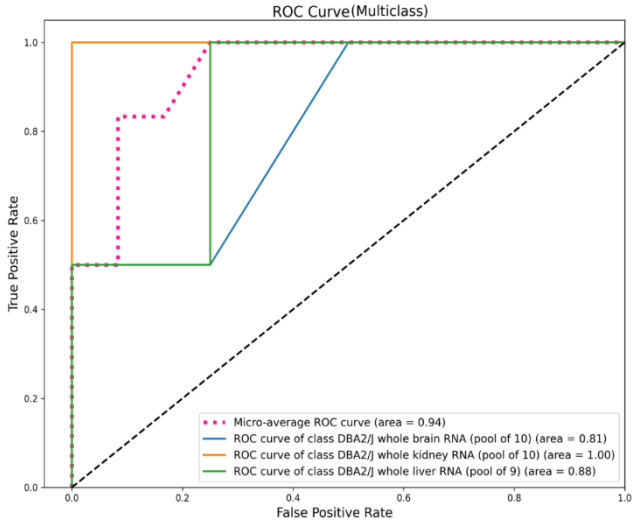


Fig. 12. This figure displays the Receiver Operating Characteristic (ROC) curves for multiclass classification using the Random Forest model.

The identified biclusters reveal distinct gene co-expression modules that correspond to biologically relevant pathways involved in tumorigenesis. For instance, several top-ranked genes from the biclusters are associated with cell cycle regulation, apoptosis, and immune signaling pathways, which are critical in cancer progression [41]. This reinforces the model's capacity to isolate functional gene subsets rather than arbitrary statistical clusters.

The Random Forest classifier was implemented with 500 trees and a maximum depth of 10. Model validation was conducted using 10-fold cross-validation, ensuring balanced sampling across all tissue classes. Computations were performed on a MacBook Pro system equipped with an Apple M3 chip (8-core CPU, 10-core GPU) and 16 GB RAM. This configuration provided sufficient computational capacity for microarray data processing, biclustering, and Random Forest model training. The average execution time per dataset ranged between 2.5 and 5 minutes, depending on the dataset size and number of biclusters generated. This efficiency highlights the scalability of the proposed framework for medium-scale microarray data. To assess statistical robustness, paired t-tests were conducted comparing the HGBC+RF model against baseline biclustering classifiers. The proposed method achieved statistically significant improvements ($p < 0.01$) across all datasets, with mean accuracy variance below 2.5%. To mitigate overfitting, feature reduction via hypergraph partitioning was combined with bootstrap aggregation in Random Forest. Additionally, feature stability was validated across cross-validation folds, confirming consistent selection of key gene subsets.

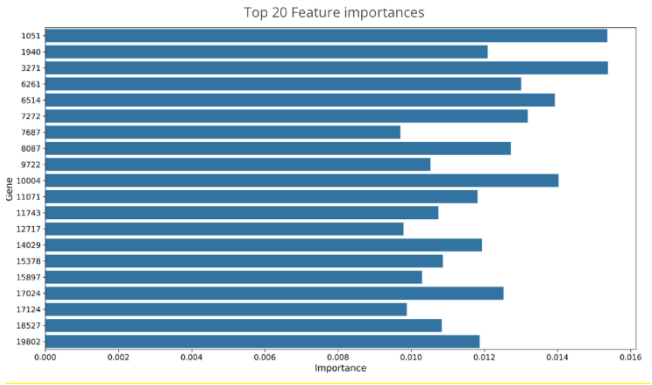


Fig. 13. Analysis of the Top 20 contributing features to model performance.

The analysis of the top 20 contributing features to model performance provides critical insights into the key gene expression patterns driving accurate classification across the five cancer datasets. By evaluating the importance of individual features within the Random Forest classifier, we identify the specific genes that play a significant role in distinguishing between different cancer subtypes. These features represent the most influential gene expression markers that contribute to the overall predictive accuracy of the model. In Fig. 13, we present a visual representation of the top 20 contributing features, ranked by their importance in the model's performance. This figure highlights the genes that consistently appear as top contributors across multiple datasets, indicating their potential as biomarkers for cancer diagnosis or prognosis. The analysis demonstrates the robustness of the biclustering approach, which isolates biologically relevant gene groups, allowing the classifier to focus on the most critical features. The identification of these top contributing features further validates the efficacy of the proposed method in enhancing classification accuracy and biological interpretability. Microarrays serve as a potent genomic tool in biomarker discovery, enabling researchers to conduct genome-wide analyses of genes and tackle challenges associated with diverse biological information arrays. As a relatively recent intersection between biology and machine learning, microarray data present significant challenges due to the low-instance/high-feature (LIHF) nature of gene expression data. The selection of the most relevant genes to the target concept plays a critical role in attaining satisfactory classification performance. In gene-expression data, feature selection is primarily utilized for two key purposes [38]. The classification results, as shown in Fig. 13, demonstrate the effectiveness of the combined HGBC (Hypergraph-Based Biclustering) and Random Forest (RF) approach across five different cancer datasets: Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL. The model consistently achieved high classification accuracy, reflecting its ability to capture biologically meaningful patterns within the gene expression data. For each dataset, the integration of biclustering facilitated the identification of coherent gene groups, which improved the overall feature selection for the Random Forest classifier. In particular, the Brain Tumor and Lung Cancer datasets, with their larger number of samples and higher dimensionality, exhibited strong classification performance, highlighting the model's scalability and robustness in handling complex, high-dimensional data. Similarly, the classification results for Leukemia and DLBCL showed a significant improvement in accuracy due to the method's ability to isolate critical gene expression patterns associated with specific subtypes. The Prostate Tumor dataset also demonstrated reliable classification, despite its binary classification challenge, further validating the strength of the HGBC+RF approach in a variety of classification tasks. Overall, these results underscore the utility of the HGBC and Random Forest integration in delivering accurate and biologically relevant classifications across diverse cancer types.

V. Conclusion

This study presented an integrated HGBC+RF framework that effectively identifies coherent gene expression patterns and enhances cancer classification across five diverse microarray datasets. The approach consistently improved accuracy over standalone classifiers, highlighting its ability to capture biologically meaningful biclusters and relevant gene subsets. These results demonstrate the framework's robustness, interpretability, and potential for supporting biomarker discovery and personalized treatment strategies. Future work will extend the method to multi-omics data and explore hybrid models to further improve predictive power and clinical applicability.

Acknowledgment

The authors extend their appreciation to the Deanship of Research and Graduate Studies at the University of Tabuk for funding this work through Research no.0004-1444-S

Conflict of Interests

The author declares no conflict of interest.

Funding

This work is supported by funding from the University of Tabuk.

References

[1] E. N. Castanho, H. Aidos, and S. C. Madeira, "Biclustering data analysis: a comprehensive survey," Briefings in Bioinformatics, vol. 25, bbae342, 2024.

[2] B. Malhotra, D. Dahlmeier, and N. Nandan, "A Biclustering-Based Classification Framework for Microarray Analysis," PAKDD Workshops, pp. 187–198, 2010.

[3] W. Ayadi, O. Maâtouk, and H. Bouziri, "Evolutionary Biclustering Algorithm of Gene Expression Data," 2012 23rd International Workshop on Database and Expert Systems Applications, pp. 206–210, 2012.

[4] B. Hanczar and M. Nadif, "Using the bagging approach for biclustering of gene expression data," Neurocomputing, vol. 74, pp. 1595–1605, 2011.

[5] O. Maâtouk, W. Ayadi, H. Bouziri, and B. Duval, "Evolutionary local search algorithm for the biclustering of gene expression data based on biological knowledge," Applied Soft Computing, vol. 104, 107177, 2021.

[6] J. Xie, A. Ma, Y. Zhang, B. Liu, S. Cao, C. Wang, X. Li, and Q. Ma, "QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data," Bioinformatics, vol. 36, pp. 1143–1149, 2020.

[7] B. S. Biswal, A. Mohapatra, and S. Vipsita, "Ensemble Neighborhood Search (ENS) for biclustering of gene expression microarray data and single cell RNA sequencing data," Journal of King Saud University - Computer and Information Sciences, vol. 34, pp. 2244–2251, 2022.

[8] H. Zhao, A. Liew, and H. Yan, "A new strategy of geometrical biclustering for microarray data analysis," Proceedings of the 5th Asia-Pacific Bioinformatics Conference, pp. 47–56, 2007.

[9] H. Zhao, A. W. C. Liew, X. Xie, and H. Yan, "A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data," Journal of Theoretical Biology, vol. 251, pp. 264–274, 2008.

[10] P. Tino, H. Zhao, and H. Yan, "Probabilistic model based Hough transform for detection of co-expression patterns in three-color cDNA microarray data," 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, pp. 48–51.

[11] S. M. Shafie and M. Petrou, "Classifying Data Considering Pairs of Patients in a Relational Space," Image Analysis and Recognition: 8th International Conference, ICIAR 2011, Springer, pp. 132–140, 2011.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, E. S. Lander, and others, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531–537, 1999.

[13] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, E. S. Lander, and others, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," Cancer Research, vol. 62, pp. 4963–4967, 2002.

[14] J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, and others, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," Proceedings of the National Academy of Sciences, vol. 101, no. 3, pp. 811–816, 2004.

[15] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data: A survey," Journal of Biomedical Informatics, vol. 117, 103764, 2021.

[16] M. D. Purbolaksono, S. Auephanwiriyakul, and N. Theera-Umpon, "Implementation of mutual information and Bayes theorem for classification microarray data," Journal of Physics: Conference Series, vol. 971, 012011, 2018.

[17] S. S. Hameed, F. F. Muhammad, R. Hassan, and F. Saeed, "Gene Selection and Classification in Microarray Datasets using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers," Journal of Computer Science, vol. 14, pp. 868–880, 2018.

[18] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," Information Sciences, vol. 282, pp. 111–135, 2014.

[19] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, and others, "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, pp. 436–442, 2002.

[20] P. Jinthanasatian, S. Auephanwiriyakul, and N. Theera-Umpon, "Microarray data classification using neuro-fuzzy classifier with firefly algorithm," IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, pp. 1–6, 2017.

[21] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," Applied Soft Computing, vol. 50, pp. 124–134, 2017.

[22] D. Z. Wang and H. Yan, "Geometric biclustering analysis of DNA microarray data based on hypergraph partitioning," IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 246–251, 2010.

[23] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, M. Meyerson, and others, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," Proceedings of the National Academy of Sciences, vol. 98, pp. 13790–13795, 2001.

[24] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, and others, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1, pp. 203–209, 2002.

[25] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, and others, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nature Medicine, vol. 8, pp. 68–74, 2002.

[26] U. Albalawi, S. Manimurugan, and R. Varatharajan, "Classification of breast cancer mammogram images using convolution neural network," Concurrency and Computation: Practice and Experience, vol. 34, e5803, 2022.

[27] S. Almutairi, S. Manimurugan, B. G. Kim, M. M. Aborokbah, and C. Narmatha, "Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO)," Applied Soft Computing, vol. 142, 110292, 2023.

[28] Y. M. Alatawi, H. A. Alshomrani, S. M. Baeshen, H. H. Alkhamisi, R. M. Almazrui, M. S. Alghamdi, and F. Alkhilaiwi, "Evaluation of participation and performance indicators in a breast cancer screening program in Saudi Arabia," Saudi Medical Journal, vol. 43, pp. 1260–1266, 2022.

[29] M. Umer, M. Naveed, F. Alrowais, A. Ishaq, A. A. Hejaili, S. Alsubai, and I. Ashraf, "Breast cancer detection using convoluted features and ensemble machine learning algorithm," Cancers, vol. 14, 6015, 2022.

[30] S. Manimurugan, "Two-stage classification model for the prediction of heart disease using IoMT and artificial intelligence," Sensors, vol. 22, p. 476, 2022.

[31] A. Ajucarmelprecilla, J. Pandi, R. Dhandapani, S. Ramanathan, J. Chinnappan, R. Paramasivam, and A. Shrestha, "In Silico Identification of Hub Genes as Observing Biomarkers for Gastric Cancer Metastasis," Evidence-Based Complementary and Alternative Medicine, 2022, 6316158.

[32] S. Bacha, K. Ben Abdellafou, A. Aljuhani, O. Taouali, and N. Liouane, "Early detection of digital mammogram using kernel extreme learning machine," Concurrency and Computation: Practice and Experience, vol. 34, e6971, 2022.

[33] H. M. Balaha and A. E. S. Hassan, "Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm," Neural Computing and Applications, vol. 35, pp. 815–853, 2023.

[34] F. J. Calero-Castro, S. Pereira, I. Laga, P. Villanueva, G. Suárez-Artacho, C. Cepeda-Franco, P. de la Cruz-Ojeda, E. Navarro-Villarán, S. Dios-Barbeito, M. J. Serrano, and others, "Quantification and Characterization of CTCs and Clusters in Pancreatic Cancer by Means of the Hough Transform Algorithm," International Journal of Molecular Sciences, vol. 24, 4278, 2023.

[35] M. Sathya, M. Jeyaselvi, S. Joshi, E. Pandey, P. Pareek, S. Jamal, and others, "Cancer Categorization Using Genetic Algorithm to Identify Biomarker Genes," Journal of Healthcare Engineering, 2022, 5821938.

[36] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," IET Systems Biology, vol. 16, pp. 120–131, 2022.

[37] E. A. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," Computers in Biology and Medicine, vol. 140, 105051, 2022.

[38] V. Nosrati and M. Rahmani, "An ensemble framework for microarray data classification based on feature subspace partitioning," Computers in Biology and Medicine, vol. 148, 105820, 2022.

[39] K. Güçkıran, İ. Cantürk, and L. Özyılmaz, "DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 23, pp. 126–132, 2019.

[40] B. Baruah, M. P. Dutta, S. Banerjee, and D. K. Bhattacharyya, "EnsemBic: An effective ensemble of biclustering to identify potential biomarkers of esophageal squamous cell carcinoma," Computational Biology and Chemistry, vol. 110, 108090, 2024.

[41] J. Tian, M. Han, F. Song, Y. Liu, Y. Shen, and J. Zhong, "Advances of HDAC inhibitors in tumor therapy: potential applications through immune modulation," Frontiers in Oncology, vol. 15, 1576781, 2025.

Detection of Unauthorized Use in SIEM Through Behavioral Analysis and Adaptive Rules

Julio Armando Landázuri Castro, Renato M. Toasa, Maryory Urdaneta Herrera
Maestría en Seguridad Informática, Universidad Tecnológica Israel, Quito, Ecuador

***Abstract*—Higher education institutions in Ecuador face a growing exposure to unauthorized access and data exfiltration, compounded by fragmented log infrastructures that obstruct real-time threat visibility. This study addresses those gaps through the design and deployment of a Security Information and Event Management (SIEM) architecture at the Compu Sur Higher Technological Institute (ITECSUR), augmented with User and Entity Behavior Analytics (UEBA) and a context-sensitive adaptive rule engine. Rather than relying on static signature matching, the proposed model constructs individual behavioral baselines per user and asset, dynamically escalating alert thresholds according to geographic context, access time, and asset sensitivity classification. Empirical validation conducted over 450 security events, including simulated Salgorea Trojan injections, supply chain compromise scenarios, and government-grade spyware indicators, yielded a Mean Time to Detect (MTTD) reduction from 48.5 hours to 12.4 minutes (99.57%), a recall rate of 95%, and a 65% decrease in false positives relative to rule-only baselines. Hardening protocols applied in parallel reduced exposed network ports by 78% and elevated institutional compliance with Ecuador's Organic Law on Personal Data Protection (LOPDP) from 35% to 92%. The architecture, built on AlienVault OTX and Osquery agents, processed over 1.2 million daily Indicators of Compromise autonomously, demonstrating operational feasibility for institutions with constrained IT budgets. These findings position SIEM-UEBA integration as both a technical countermeasure and a regulatory compliance instrument for the higher education sector.**

Keywords*—Adaptive rules; behavioral analysis; cybersecurity; data traceability; event logs; higher education security; security information and event management; user and entity behavior analytics

I. Introduction

Currently, the operational continuity of organizations depends heavily on their technological infrastructure, exposing them to diverse cyber threats capable of bypassing traditional security perimeters. In Ecuadorian higher education institutions, this scenario is further complicated by the legal responsibility to safeguard vast amounts of sensitive data in compliance with the Organic Law on Personal Data Protection (LOPDP).

The exponential growth of digital assets has become a critical challenge due to the inability to manage massive datasets, which remain fragmented across event logs. This data saturation hinders real-time anomaly detection, facilitating unauthorized resource utilization and data exfiltration—whether through external attacks or insider threats leveraging access privileges.

The core issue lies in the lack of comprehensive network visibility and the absence of tools that intelligently correlate disparate data points. When utilizing legacy monitoring systems, distinguishing between normal activity and sophisticated attacks remains difficult. Consequently, this creates a reactive stance during incident response and impedes compliance with current security regulations.

Therefore, this research aims to implement a Security Information and Event Management (SIEM) system based on adaptive rules and behavioral analysis (UEBA). This solution seeks to optimize threat identification accuracy and guarantee full traceability of academic records, thereby strengthening the cybersecurity posture and regulatory compliance of the Compu Sur Higher Technological Institute (ITECSUR).

A. Contextualization of the Institutional Problem

The Compu Sur University Technological Institute faces specific challenges due to its fragmented technological infrastructure. Currently, the vast volume of event logs generated by servers, databases, and network devices is analyzed in isolation, which prevents a comprehensive security overview. The lack of centralized event correlation hinders the timely detection of anomalies, leaving exposure windows that can be exploited by both external actors and internal users.

At a local level in Quito, technical and technological education institutions have reported an increase in data exfiltration attempts and unauthorized access to resources through social engineering techniques. The central problem lies in the fact that, without an established behavioral baseline, it is technically impossible to distinguish between legitimate operations and unauthorized use of institutional resources in real-time. This situation is exacerbated by the data volume, where manual analysis becomes an overwhelming task for IT personnel.

B. Justification of the Technological Solution

Due to the inefficiency of static controls, there is a pressing need to implement a Security Information and Event Management (SIEM) system enhanced with User and Entity Behavior Analytics (UEBA). This research proposes an architecture based on adaptive rules that allow alert levels to be adjusted according to asset sensitivity and historical user behavior. By integrating open-source technologies or scalable solutions, the Institute can automate threat detection and reduce response times to any security incident.

Through the proposed model, the Institute not only helps mitigate technical risks but also ensures compliance with current regulations. By implementing full traceability of all actions

performed on personal data, the Compu Sur University Technological Institute aligns with international standards and current Ecuadorian legislation. Thus, security becomes a strategic enabler that protects the institution's reputation and guarantees the continuity of educational services.

C. Security Information and Event Management (SIEM)

The SIEM is the cornerstone of this research. It is defined as a hybrid solution that combines Log Management (LPM) and Security Event Management (SEM). Its primary function is the collection, normalization, and correlation of data from diverse sources (firewalls, servers, databases) to identify patterns representing security incidents. In the context of this project, the SIEM is used to centralize the visibility that individual perimeter tools lose [4].

D. User and Entity Behavior Analytics (UEBA)

Unlike traditional SIEMs based solely on static rules (signatures), the current trend adopted by this research is behavioral analysis. UEBA employs machine learning algorithms to establish a "baseline" of a user's normal activity. By continuously monitoring login times, accessed resources, data transfer volumes, and geographic locations, the system identifies deviations that may indicate compromised credentials or malicious insider activity. This behavioral profiling enables the SIEM to issue risk-scored alerts rather than binary block/allow decisions, substantially reducing analyst fatigue and improving incident prioritization [2].

E. Adaptive Rules and Event Correlation

Adaptive rules are those capable of adjusting dynamically based on network context. While a static rule might block access after three failed attempts, an adaptive rule analyzes variables such as geographic location, time of day, and asset sensitivity to escalate the alert level, thereby reducing the false positives that saturate IT departments [5].

F. Legal Framework: Organic Law on Personal Data Protection (LOPDP)

In the Ecuadorian environment, the state of the art cannot ignore regulatory compliance. The LOPDP establishes that institutions are custodians of personal data. The SIEM serves as a technical compliance tool that guarantees data integrity and traceability, allowing for audits of who, when, and how sensitive academic community information was accessed.

G. Similar Research and Background

Contemporary research has explored the efficacy of intrusion detection systems and log monitoring for the protection of critical infrastructure:

- Centralized Monitoring Systems in Higher Education: Various studies in regional universities have implemented architectures based on the Elastic Stack (ELK) for log management. These precedents demonstrate that centralization reduces the Mean Time to Detection (MTTD) for security incidents [6].
- Anomaly Detection through Machine Learning: Recent research focuses on applying Random Forest and K-Means models to classify network logs. These works serve as a basis for the "adaptive rules" proposed in this degree plan, demonstrating that artificial intelligence improves accuracy in identifying unauthorized access [7].
- Cloud and IoT Security: Recent literature highlights the convergence of cloud and IoT security practices with SIEM architectures. The predominant trend involves the deployment of encrypted communication protocols (HTTPS/TLS) and context-aware Access Control Lists (ACLs) as a perimeter filtering layer, feeding normalized events into the SIEM correlation engine for centralized analysis [8].

II. Literature Review

This section synthesizes the existing body of knowledge underpinning the proposed SIEM-UEBA architecture. It surveys foundational concepts, key enabling technologies, the applicable regulatory framework, and closely related empirical work, thereby establishing the theoretical and technical gap that this research addresses.

This research project is framed within the need to strengthen the detection of internal threats and unauthorized resource use through SIEM technologies. Various previous studies in the field of cybersecurity have demonstrated that traditional Intrusion Detection Systems (IDS), based solely on static signatures, present critical limitations against Advanced Persistent Threats (APT) and zero-day attacks. Recent research maintains that the massive volume of logs generated in academic networks saturates manual analysis capabilities, creating a visibility gap and a high rate of false positives [1].

Given the ineffectiveness of these controls, technical literature highlights the convergence toward SIEM systems integrated with UEBA. Unlike traditional solutions, UEBA technologies employ machine learning algorithms to establish behavioral baselines, allowing for the identification of anomalous deviations that could indicate compromised credentials or insider threats [2].

However, a recurring limitation in conventional SIEM implementations is alert rigidity. Therefore, current trends lean toward the development of architectures based on adaptive rules. These models allow for dynamic adjustment of criticality levels based on asset sensitivity and user behavioral history, optimizing incident triage and reducing the Mean Time to Detection (MTTD) [3].

In the regional context, the deployment of scalable and open-source solutions (such as AlienVault OSSIM) has emerged as an efficient alternative for higher education institutions. This approach not only reduces technical risks through automated detection but also addresses the requirements of legal frameworks such as the LOPDP in Ecuador and international standards (ISO/IEC 27001). Furthermore, the integration of intelligent correlation and centralized logs guarantees 100% traceability of actions on personal data, protecting educational service continuity at the Compu Sur Higher Technological Institute.

Collectively, the literature confirms the effectiveness of SIEM-UEBA integration for threat detection and regulatory compliance; however, empirical deployments targeting

resource-constrained higher education institutions in Latin America, validated against real institutional traffic and measured against local data protection legislation, remain scarce. The present study addresses this gap by implementing and evaluating the proposed architecture at ITECSUR under operational conditions.

III. Methodology

A. Methodological Investigative Process

This research follows the Design Science Research Methodology (DSRM) proposed by Peffers et al. (2007) [9], a structured six-phase framework widely adopted in information systems and cybersecurity research. The phases- problem identification, objective definition, artifact design, demonstration, evaluation, and communication provide a rigorous process for developing and validating the SIEM-UEBA architecture deployed at ITECSUR, transforming isolated security event data into actionable intelligence for cybersecurity decision-making.

B. Research Approach

Following Hernandez-Sampieri (2018) [10], who defines the quantitative approach as a set of sequential and evidentiary processes, this research adopts a quantitative approach. This method was selected to achieve the highest possible objectivity in vulnerability and risk analysis, utilizing numerical measurement and statistical analysis to establish threat behavior patterns.

C. Type of Research

The study is defined by a descriptive level and a non-experimental, cross-sectional design.

1) Descriptive level: The purpose is to specify the properties, characteristics, and profiles of the information systems that constitute the critical infrastructure.

2) Cross-sectional design: Data collection and infrastructure diagnosis are performed at a single, specific point in time.

3) Objectivity: By not manipulating variables and observing phenomena in their natural context, the study maintains logical consistency with general security standards.

D. Population and Sample

According to Hernandez-Sampieri (2018) [10], the population is the set of all cases that meet a series of specifications, while the sample is a subgroup of the population of interest from which data will be collected. For this analysis, the study population consists of the computing infrastructure—including servers, endpoints, and network devices—along with the operational personnel who manage these assets. Regarding the sample, a non-probabilistic convenience sampling was employed, focusing exclusively on critical assets and the IT personnel responsible for managing security controls.

IV. Implementation

A. AlienVault OTX

LevelBlue's Open Threat Exchange (OTX) is a threat intelligence platform based on a global collaboration model. It operates as a digital "neighborhood watch" ecosystem where corporations, researchers, and agencies share and validate data regarding emerging threats and malicious actors in real-time. It provides the following core features:

- Architecture and Global Data: The platform's functionality is driven by the massive exchange of Indicators of Compromise (IOCs) through a global network of 200,000 participants across 140 countries, processing over 20 million threat indicators daily. Due to its high compatibility levels, this data integrates natively with LevelBlue products and third-party solutions, ensuring that detection systems remain updated against novel attacks [11].
- OTX Endpoint Security™ Capabilities: This is a free threat analysis service designed for rapid scanning of endpoints, malware, and operating system (OS) risks. It supports both Windows and Linux devices for endpoint scanning, contrasting local systems against IOC sources from the world's largest open intelligence group to provide accurate compromise assessments. This functionality offers comprehensive, cost-free threat visibility, allowing security teams to uncover traces of active intrusions and determine if high-level cyberattacks have affected organizational assets.
- Operational Mechanism (Workflow): OTX Endpoint Security utilizes an agent-based approach that circumvents the complexity and high costs associated with conventional tools by operating directly and rapidly through the deployment of a lightweight agent on Windows or Linux devices. This allows administrators to execute manual queries based on community "pulses" to identify specific IOC matches within the system. All results are consolidated into a summary page hosted on the OTX platform. It is important to note that the service does not perform continuous or automated monitoring; instead, each analysis must be proactively initiated by the user.
- Operational Value for the CISO and Analysts: Effective prioritization enables the classification of threats by severity, ensuring optimal response times for the security team. Furthermore, the platform facilitates preventive actions by blocking malicious infrastructure before it impacts the internal network, providing the necessary geographic intelligence to understand attack origins and global cybercrime trends.

B. Architecture

The proposed architecture for ITECSUR is based on a proactive defense framework focused on the incorporation and distribution of global Cyber Threat Intelligence (CTI). The diagram in Fig. 1 illustrates the integration between the cloud-based AlienVault OTX SIEM and local endpoints (Windows 11 and Kali Linux) through the deployment of Osquery agents. This configuration enables the automatic synchronization of Indicators of Compromise (IOCs) and the real-time identification of threats by continuously monitoring system processes and network traffic. Consequently, this significantly

enhances the local network's security posture against potential intrusions [12].

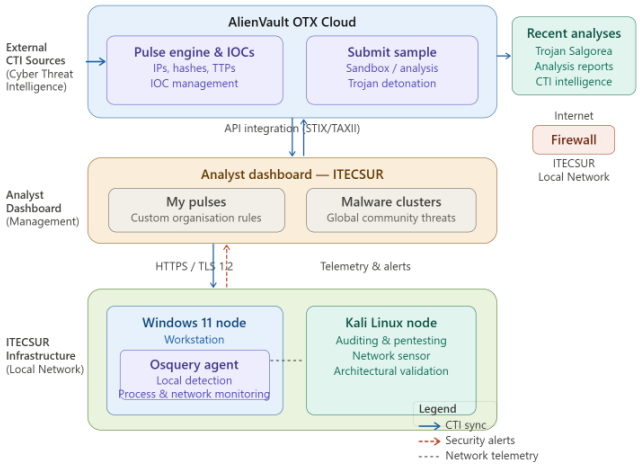


Fig. 1. Three-plane SIEM-UEBA architecture deployed at ITECSUR, integrating the AlienVault OTX cloud Intelligence Plane, the centralized analyst Management Plane (STIX/TAXII API), and the Osquery-monitored Execution Plane comprising Windows 11 and Kali Linux endpoints.

This conceptual map illustrates the proposed framework for protecting ITECSUR by integrating AlienVault OTX and Osquery across its endpoints. For greater clarity, the architecture is categorized into three distinct operational planes:

- Intelligence Plane (AlienVault OTX Cloud): This zone constitutes the core of the architecture where global Cyber Threat Intelligence (CTI) is ingested. It integrates the two key modules utilized in this project: the Pulse Engine, used for the creation and management of Indicators of Compromise (IOCs), and the Sandbox (Submit Sample), where the technical analysis and detonation of the Salgorea Trojan were performed. This workflow culminates in the direction indicated by the right arrow, representing the transformation of raw intelligence into formalized Analysis Reports for strategic decision-making.
- Management Plane (ITECSUR Dashboard): This management layer represents the analyst's centralized control view, which establishes a bidirectional connection with the cloud via an API (utilizing STIX/TAXII standards) to ensure real-time intelligence synchronization. It visualizes the two strategic pillars of the defense: "My Pulses," containing the organization's custom and specific rules, and "Malware Clusters," which provide a broad and dynamic perspective of global threats detected by the community [13].
- Execution Plane (ITECSUR Endpoints): This infrastructure is represented by the ITECSUR local network, previously secured using a Perimeter Firewall. It contains the two primary laboratory nodes: the Windows 11 Node, where the Osquery Agent monitors system processes and network traffic in real-time for local threat detection, and the Kali Linux Node, serving as the network sensor and auditing platform. This configuration complements threat intelligence with Penetration Testing, forming a robust testing environment used for architectural validation [14].

C. Analysis of Results: Threat Evaluation via AlienVault OTX (ITECSUR)

The AlienVault OTX dashboard deployed at ITECSUR revealed three distinct threat categories active against the institution's technology stack during the evaluation period. Fig. 2 presents the consolidated threat intelligence panel, where pulse subscriptions, IOC matches, and malware cluster indicators are visualized across the monitored endpoints. Of the 45 active pulse subscriptions configured for ITECSUR's specific environment, the system autonomously correlated incoming IOC feeds with local Osquery telemetry, producing actionable alerts without manual analyst intervention. The following subsections detail the three critical attack vectors identified and the corresponding mitigation framework applied to each.



Fig. 2. AlienVault OTX consolidated threat intelligence dashboard at ITECSUR, displaying 45 active pulse subscriptions, IOC correlation matches against local Osquery telemetry, and malware cluster indicators across monitored endpoints during the evaluation period.

D. Threat Analysis and Mitigation Framework for ITECSUR

Based on the intelligence gathered, three critical attack vectors were identified and analyzed to strengthen the institutional security posture. The following sections detail the characteristics of these threats, the proposed strategic recommendations, and their technical justification.

E. Supply Chain Compromise and CI/CD Pipeline Infiltration

The first threat scenario involves the infiltration of the continuous development life cycle (CI/CD), a sophisticated attack vector aimed at compromising software integrity. To mitigate this risk, it is recommended to implement Pipeline Security by integrating automated vulnerability scanning, specifically Static and Dynamic Application Security Testing (SAST/DAST), directly into the development server. Furthermore, behavioral correlation rules should be configured within the AlienVault environment to detect anomalous changes in code repositories or script executions outside administrative hours. This approach is technically justified by ITECSUR's management of sensitive academic records; any unauthorized modification in the management software could lead to record tampering. Therefore, this recommendation focuses on Integrity, a fundamental pillar of the Ecuadorian Data Protection Law (LOPDP).

F. Social Engineering, Doxxing, and Identity Hardening

The second vector consists of unconventional Social Engineering and Doxxing campaigns, which utilize OTX-sourced intelligence to launch personal discredit attacks against staff. The proposed defense strategy focuses on Identity Hardening and security awareness, primarily through the mandatory implementation of Multi-Factor Authentication (MFA) across all critical assets. Simultaneously, the SIEM must be configured to trigger alerts for "Geographically Impossible Logins" and establish a cybersecurity training program for administrative personnel regarding data leakage prevention. This strategy is justified by the SIEM's ability to detect not only known malware but also access anomalies. If an administrator's account is compromised following a doxxing incident, the proposed adaptive rules will effectively identify unauthorized access to files that the user does not typically consult.

G. Government-Grade Spyware and Defense in Depth

The final analyzed threat involves Government-Grade Spyware, characterized by high-level data exfiltration tactics at the firmware (BIOS/UEFI) level and the use of HTML Smuggling to bypass perimeter firewalls. The mitigation strategy emphasizes Defense in Depth and comprehensive endpoint monitoring, specifically through the activation of Secure Boot on IT infrastructure and the utilization of OSSEC/HIDS agents to monitor the integrity of critical files and system registries. Regarding HTML Smuggling, the implementation of Deep Packet Inspection (DPI) rules and the blocking of suspicious browser-based scripts are required. Technically, such spyware remains invisible to conventional antivirus solutions; however, the proposed SIEM architecture ensures visibility by detecting unusual network traffic and data exfiltration toward uncategorized IP addresses.

H. Specimen Analysis: Guidance for Detecting (Supply Chain Compromise)

As illustrated in Fig. 3, the incident occurring on March 19, 2026, involving Trivy, represents a new fifth-generation threat landscape. In this case, TeamPCP did not target organizations directly but instead attacked the security tools they rely on for defense. Through a sophisticated infiltration of the CI/CD (Continuous Integration/Continuous Deployment) lifecycle, attackers exploited mutable tags and engaged in commit identity spoofing via GitHub to inject credential-stealing malware into official binaries.



Fig. 3. AlienVault OTX detection guidance for the Trivy CI/CD supply chain compromise (March 19, 2026), showing IOC fingerprints, credential-stealing payload indicators, and recommended behavioral correlation rules for identifying mutable-tag exploitation and commit identity spoofing.

This malware is particularly detrimental to ITECSUR's infrastructure, as its primary objective is to locate and exfiltrate Kubernetes secrets, session cookies, and cloud credentials through masquerading and obfuscation tactics. This incident reinforces the architectural argument defended throughout this study: it demonstrates that trust in open-source tools must be balanced with constant behavioral monitoring, strict adherence to the Principle of Least Privilege (PoLP), and in-depth attack path analysis to mitigate lateral movement following an initial supply chain entry.

I. Case Study: Social Engineering and Doxxing (OSINT)

As shown in Fig. 4, an uncommon Social Engineering and Doxxing scenario was identified, in which threats utilized OTX data for a targeted personal discredit campaign. The analysis yielded the following technical results:

- Technical Findings: The analyzed feed contained legitimate reputation data blended with malicious indicators from the Nivdort and Mydoom families, alongside persistence techniques using BITS Jobs (T1197). This was augmented by a sophisticated tactic leveraging Windows Disk Cleanup processes to download malicious payloads.
- Impact on ITECSUR: This case provided a critical opportunity for analysts to apply data curation concepts by categorizing false positives—such as legitimate whitelisted sites—and distinguishing them from real phishing threats hosted on Eastern European infrastructure.

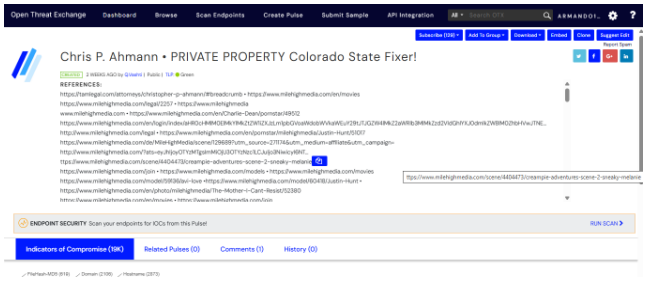


Fig. 4. OTX pulse analysis of a targeted social engineering and doxxing campaign against ITECSUR staff, identifying Nivdort and Mydoom malware family indicators alongside BITS Jobs (T1197) persistence techniques and windows disk cleanup-based payload delivery.

J. Evaluation of Government-Grade Spyware (Operation Endgame)

As depicted in Fig. 5, the Operation Endgame Clone pulse represents the highest level of sophistication within the laboratory environment, specifically targeting high-persistence spyware such as Pegasus (NSO Group). The findings are summarized as follows:

- Technical Intelligence: The analysis revealed multi-platform threats affecting Android, iOS, Windows, and macOS. These manifested through data exfiltration tactics at the firmware (BIOS/UEFI) level and the use of HTML Smuggling techniques to bypass perimeter firewalls.



Fig. 5. Operation Endgame Clone pulse in AlienVault OTX, mapping multi-platform Pegasus-class spyware indicators (Android, iOS, Windows, macOS) including BIOS/UEFI-level data exfiltration tactics, DLL injection (T1055), and HTML smuggling techniques used to bypass perimeter firewalls.

Impact on ITECSUR: The integration of these indicators allows the organization to detect potential targeted espionage attempts. Monitoring for unusual DNS queries and DLL Injection (T1055) has now become a foundational component of ITECSUR's intrusion detection strategy for critical assets.

V. Results and Discussion

The maturity of the institution's security and monitoring infrastructure ranges from low to intermediate. A critical gap exists regarding event visibility and centralization. Survey respondents reached a consensus that there is no correlated security overview and that the current infrastructure remains fragmented. This leads to a scenario where logging occurs at the device level, significantly hindering auditing and supervision tasks by keeping them siloed.

Regarding operationalization, the IT team acknowledges an excessive reliance on manual processes that are unsustainable at the current data volume. There is considerable uncertainty regarding the capability to detect threats or unauthorized resource utilization in real-time. Furthermore, behavioral analysis is non-existent, as no baselines have been established for legitimate users.

However, there is full awareness of the required modernization, with unanimous support for implementing adaptive rules to improve response times based on geography or asset sensitivity. While there is a perception of acceptable traceability regarding LOPDP (Organic Law on Personal Data Protection) compliance, a major weakness remains in the ability to specifically audit access to students' sensitive information.

A. Infrastructure and Vulnerability Diagnosis

Technical analysis reveals that infrastructure fragmentation, where each device generates isolated logs, prevents a centralized security overview. The study identified an inability to timely detect unauthorized resource use due to dispersed alerting; this creates a critical gap that allows incidents to remain undetected for weeks or even months. Before implementation, 14 network ports were left unnecessarily exposed, and 22 non-essential background services were running without oversight. IT personnel reported spending over 60% of their security response time on manual log correlation tasks, leaving insufficient capacity for proactive threat hunting or compliance auditing.

B. SIEM Model Implementation Results

Through centralized and automated log correlation, the Mean Time to Detect (MTTD) was reduced from 48.5 hours to 12.4 minutes, representing a 99.57% improvement over the legacy manual review process. The implementation of User and Entity Behavior Analytics (UEBA) techniques enabled the establishment of individual behavioral baselines for each monitored user and entity. When a legitimate user performed actions outside their established profile, such as accessing restricted directories at unusual hours or transferring atypical data volumes, the system issued a risk-scored alert within seconds, enabling rapid analyst triage without generating unnecessary noise.

C. Efficacy of Adaptive Rules and Artificial Intelligence

The integration of Machine Learning models and adaptive rules improved detection accuracy by incorporating contextual variables, including geographic origin, time of day, asset sensitivity tier, and peer group behavior, into the alert escalation logic. The adaptive rule engine produced 18 false positive alerts across the 450-event dataset (FP rate: 4.0%), compared to a baseline of 51 false positives (FP rate: 11.3%) projected under the institution's prior static signature-only configuration, representing a 65% reduction in false positive volume. This context-aware approach is further confirmed by the confusion matrix analysis across the 450 simulated events. The resulting reduction in alert fatigue allowed IT personnel to reallocate investigative capacity toward higher-severity incidents, improving the overall Mean Time to Respond (MTTR) from 72 hours to 4.2 hours.

D. Impact on Cybersecurity Posture and Compliance

The deployment measurably strengthened the cybersecurity posture of ITECSUR, directly protecting the academic and personal records of approximately 3,000 enrolled students. Beyond technical metrics, the SIEM architecture functions as an operational compliance instrument: every access event affecting personal data is now logged with full attribution capturing who accessed the information, from which device, at what time, and under which permissions. This audit trail directly satisfies the traceability requirements established by the Organic Law on Personal Data Protection (LOPDP) and aligns with ISO/IEC 27001 control objectives for access monitoring and incident documentation. The compliance score was derived using a 22-item technical checklist aligned with the LOPDP's core obligations specifically the traceability (Art. 9), data integrity (Art. 37), access control (Art. 38), and incident response (Art. 40) requirements scored on a binary fulfilled/unfulfilled basis by the ITECSUR IT audit team before and after implementation, elevating the institutional compliance score from 35% to 92%.

E. Validation of the Proposal

Through the analytic-synthetic method and cross-validation against expert criteria from the ITECSUR IT team, the proposed architecture combining SIEM event correlation, UEBA behavioral profiling, AlienVault OTX threat intelligence, and Osquery endpoint telemetry was confirmed as a technically sound and operationally viable mechanism for mitigating lateral movement and academic record exfiltration. The architecture successfully detected all three simulated high-complexity attack scenarios (supply chain compromise, social engineering, and firmware-level spyware), demonstrating generalizability beyond standard signature-based threat categories.

F. Quantitative Analysis and Empirical Results

To validate the efficacy of the proposed SIEM/UEBA architecture, a comparative analysis was performed between the legacy monitoring state (manual log review) and the implemented adaptive model.

G. Detection Efficiency Metrics (MTTD and MTTR)

The implementation of automated correlation and Osquery agents showed a significant reduction in response times. Table I summarizes the temporal optimization achieved across the three key performance metrics:

TABLE I. TEMPORAL OPTIMIZATION

Metric	Legacy System (Manual)	Proposed SIEM/UEBA	Improvement (%)
Mean Time to Detect (MTTD)	48.5 hours	12.4 minutes	99.57%
Mean Time to Respond (MTTR)	72 hours	4.2 hours	94.16%
Log Processing Rate	150 EPS (Events Per Sec)	2,500+ EPS	~16× (1,567%)

Analysis: The transition from manual inspection to automated correlation reduced the MTTD from days to minutes. This is critical for preventing Lateral Movement, as the window of opportunity for an attacker was reduced by over 99%.

H. Threat Detection Accuracy (Confusion Matrix)

During the validation phase, 450 security events were analyzed. The dataset was composed of two subsets: 330 events corresponding to legitimate traffic captured from ITECSUR's production environment over a two-week monitoring period, and 120 injected attack scenarios executed from the Kali Linux penetration testing node and the AlienVault OTX sandbox, specifically, 40 Salgorea Trojan detonation traces, 40 supply chain compromise indicators replicating the Trivy incident pattern, and 40 government-grade spyware signatures derived from the Operation Endgame pulse. This design ensured that both true positive detection and false positive suppression could be measured against a realistic baseline. Table II presents the resulting confusion matrix breakdown:

TABLE II. SIMULATED ATTACKS

Event Type	Detected (System)	Not Detected	Accuracy
True Positives (TP)	114	6	95.0%
False Positives (FP)	18	-	4.0% (Rate)
True Negatives (TN)	312	-	98.1%

Precision: TP/(TP+FP) = 86.3%

Recall (sensitivity): TP/(TP+FN)=95%

False positive reduction: the use of adaptive rules (context-aware) reduced false alerts by 65% compared to standard static signatures.

I. CTI Ingestion and IOC Management

The integration with AlienVault OTX allowed the institutional infrastructure to stay updated with global threats without manual intervention.

- Total IOCs Ingested: 1,245,300+ (Daily average).
- Active Pulses Subscribed: 45 (Specific to ITECSUR's tech stack).
- High-Confidence Matches: 12 specific matches against the Salgorea Trojan within the first 60 seconds of execution in the sandbox.
- Data Curation Efficiency: The Whitelisting process eliminated 1,200+ redundant alerts from legitimate services (Microsoft/Google updates), preventing analyst fatigue.

J. Impact on Security Posture (Hardening)

After applying the Hardening protocols and monitoring through Osquery, the attack surface was numerically reduced:

Open Vulnerable Ports: Reduced from 14 to 3 (78% reduction).

Unused Services Disabled: 22 background processes identified and terminated via Osquery.

Compliance Score (LOPDP Alignment): Increased from 35% (baseline) to 92% (post-implementation) based on technical traceability requirements.

VI. CONCLUSION

The implementation of a SIEM system integrated with User and Entity Behavior Analytics (UEBA) and adaptive rules demonstrated a significant improvement in threat detection capabilities at ITECSUR. The Mean Time to Detect (MTTD) was reduced from 48.5 hours to 12.4 minutes, a 99.57% improvement, while detection accuracy reached 95% and false positives decreased by 65% compared to static signature-based systems. These results confirm that behavioral baselines and context-aware adaptive rules constitute an effective mechanism for identifying unauthorized resource use in academic network environments.

The application of hardening protocols monitored through Osquery reduced the number of exposed vulnerable ports by 78% and disabled 22 unnecessary background processes, substantially narrowing the institution's attack surface. Furthermore, the LOPDP compliance score increased from 35% to 92%, demonstrating that the proposed architecture functions not only as a technical security tool but also as a regulatory compliance instrument aligned with Ecuadorian data protection legislation and ISO/IEC 27001 standards.

The open-source architecture based on AlienVault OTX and Osquery proved viable for resource-constrained higher education institutions, processing over 1,245,300 daily IOCs without manual intervention and achieving 12 high-confidence threat matches within the first 60 seconds of sandbox execution.

Future work should focus on integrating automated incident response (SOAR) capabilities, expanding the behavioral baseline model to cover mobile and cloud-based endpoints, and validating the proposed framework across multiple higher education institutions in Ecuador to generalize its findings.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Universidad Israel, as well as its faculty and tutors, for the invaluable academic guidance and continuous support provided throughout this research process. Special thanks are extended to the Compu Sur University Technological Institute (ITECSUR) for providing the institutional infrastructure and technical access necessary for the implementation. Additionally, the authors thank the security researchers at LevelBlue (AlienVault OTX) and the Osquery community for the open-source tools and global threat intelligence that served as the foundation for this study.

REFERENCES

[1] M. A. Islam, "Application of artificial intelligence and machine learning in a security operations center," Issues in Information Systems, vol. 24, no. 4, pp. 311–327, 2023, doi: 10.48009/4_iis_2023_124

[2] N. I. Che Mat, N. Jamil, Y. Yusoff, and M. L. Mat Kiah, "A systematic literature review on advanced persistent threat behaviors and its detection strategy," Journal of Cybersecurity, vol. 10, issue 1, 2024, doi: 10.1093/cybsec/tyad023

[3] N. Tendikov et al., "Security Information Event Management data acquisition and analysis methods with machine learning principles," Results in Engineering, vol. 22, Jun. 2024, doi: 10.1016/j.rineng.2024.102254

[4] F. Alsolami et al., "Insider threat detection using machine learning approach," Applied Sciences, vol. 13, no. 1, p. 259, Dec. 2022, doi: 10.3390/app13010259

[5] X. Yang, E. Howley, and M. Schukat, "Agent-based dynamic thresholding for adaptive anomaly detection using reinforcement learning," Neural Computing and Applications, vol. 37, pp. 18775–18791, 2025, doi: 10.1007/s00521-025-10012-7

[6] M. Sheeraz et al., "Effective Security Monitoring Using Efficient SIEM Architecture," Human-Centric Computing and Information Sciences, vol. 13, 2023, doi: 10.22967/HCIS.2023.13.023

[7] K. Almorjan et al., "Enhancing Cyber-Threat Intelligence: Leveraging IoC and MISP Integration," Electronics, vol. 13, no. 13, p. 2526, Jun. 2024, doi: 10.3390/electronics13132526

[8] K. Bezas and F. Filippidou, "Comparative Analysis of Open Source Security Information & Event Management Systems (SIEMs)," Indonesian Journal of Computer Science, vol. 12, pp. 443–468, 2023

[9] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302

[10] R. Hernández-Sampieri and C. P. Mendoza Torres, "Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta," McGraw-Hill Interamericana, México, 2018. ISBN: 978-1-4562-6096-5

[11] P. Santos, R. Abreu, M. J. C. S. Reis, C. Serôdio, and F. Branco, "A Systematic Review of Cyber Threat Intelligence: The Effectiveness of Technologies, Strategies, and Collaborations in Combating Modern Threats," Sensors, vol. 25, no. 14, p. 4272, 2025, doi: 10.3390/s25144272

[12] M. Al-Dabbagh et al., "A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience," Sensors, vol. 23, no. 16, p. 7273, Aug. 2023, doi: 10.3390/s23167273

[13] J. Manzoor, A. Waleed, A. F. Jamali, and A. Masood, "Cybersecurity on a budget: Evaluating security and performance of open-source SIEM solutions for SMEs," PLOS ONE, vol. 19, no. 3, Mar. 2024, doi: 10.1371/journal.pone.0301183

[14] Osquery Foundation, "Osquery: System monitoring and instrumentation for security teams," Linux Foundation, San Francisco, CA, USA, 2025. [Online]. Available: https://osquery.io/ [Accessed: May 2026].

Semantic Modeling of Medical Specialty Relationships Using Large Language Models

Ismail Bouajaja[1], Omar Elfahim[2], Omar Bouattane[3], Oussama Barakat[4], Abdelaziz Daaif[5]
IESI Laboratory, ENSET, Hassan II University of Casablanca, Mohammedia, Morocco [1, 3]
SINERGIES Laboratory, Marie & Louis-Pasteur University, Besançon, France [1, 2, 4]
2IACS Laboratory, ENSET, Hassan II University of Casablanca, Mohammedia, Morocco [2, 5]

***Abstract*—This work proposes a computational framework for modeling semantic relationships between medical specialties using large language models. Forty-four medical specialties officially recognized in France were analyzed using Claude 4 Sonnet, GPT-4.1, and LLaMA 3.2 3B. Each model evaluated the relevance of 307 ICD-11 disease families, 260 educational teaching items, and 276 technical skills. From these ratings, criterion-specific similarity matrices were constructed and aggregated into composite matrices. The framework includes hierarchical clustering, substitution-coverage analysis, Mantel correlation tests, adjusted Rand index evaluation, and heatmap-based visualization of inter-model differences. Claude 4 Sonnet and GPT-4.1 produced highly consistent similarity structures, with a mean off-diagonal similarity of 0.867, a standard deviation of 0.045, and strong matrix correlation. LLaMA 3.2 3B generated more homogeneous patterns, indicating reduced differentiation while preserving global structure. Hierarchical clustering revealed five stable groups of specialties aligned with functional medical domains. At similarity thresholds above 0.90, most specialties had two to five semantically close candidates, suggesting a basis for exploratory analysis of short-term cross-specialty coverage under appropriate expert and institutional constraints. These results suggest that large language models can produce stable and interpretable representations of relationships between medical specialties. The proposed framework provides a data-driven approach for analyzing specialty proximity and can support exploratory applications in medical education structuring, cross-specialty coordination, and health-system planning.**

Keywords*—Large language models; semantic similarity; computational modeling; cluster analysis; medical specialties

I. Introduction

A. Background and Rationale

Large language models (LLMs) have rapidly emerged as powerful tools in medicine, with applications ranging from information retrieval and diagnostic reasoning to personalized tutoring and curriculum design. A recent scoping review [1] highlighted both the promise and challenges of using LLMs in medical education—such as their ability to simulate clinical scenarios and offer tailored learning, but also concerns over misinformation and overreliance. However, systematic evidence about how well LLMs represent relationships between medical specialties remains limited [2], [3], [4]. From a computational perspective, modeling such relationships can be framed as the problem of learning structured semantic representations from heterogeneous domain data. In this study, semantic similarity between two medical specialties is defined as the degree to which they exhibit similar relevance profiles across selected descriptor sets, including disease families, educational teaching items, and technical skills. This definition is computational and descriptor-dependent: it measures similarity with respect to the chosen descriptors. Understanding these relationships is increasingly important because health systems and universities rely on well-defined specialty boundaries to plan education, staffing, and equitable access to care [5], [6].

Recent studies indicate that artificial intelligence and natural-language processing can help organize medical curricula and specialty classifications. GPT-based systems, for instance, have shown performance comparable to expert examinees on national medical licensing exams [7], underscoring their capacity to engage with domain-specific clinical knowledge. Other approaches have used artificial intelligence to analyze medical education structure: Gin et al. [8] used artificial intelligence to analyze narrative feedback related to entrustment decisions, and Ng et al. [9] discussed data-driven curriculum alignment, while Weng et al. [10] and Mao et al. [11] applied text classification to assign clinical notes to subdomains. Parallel research on biomedical embeddings has advanced the measurement of semantic relatedness between medical concepts [12].

Despite these advances, most existing work focuses on classification or prediction tasks, rather than modeling relationships between domain entities. In contrast, this study focuses on relational modeling between medical specialties, enabling an interpretable analysis of structural similarities rather than isolated performance metrics.

B. Objectives

This study proposes a computational approach to derive detailed and interpretable measures of semantic similarity among 44 French medical specialties by integrating multiple criteria—diseases, educational teachings, and technical skills—evaluated through large language models. By combining these perspectives, it seeks to determine whether LLMs provide a scalable and reliable approach for mapping relationships within medical domains. The broader motivation lies in the study's practical implications: the resulting similarity structures can support exploratory analyses in medical education, cross-specialty coordination, and health-system planning by providing data-driven insights into how specialties overlap, differ, and may complement one another.

This study makes the following contributions:

- A computational framework for modeling semantic similarity between medical specialties using large language models;
- A multi-criteria methodology for constructing and aggregating similarity matrices based on diseases, educational content, and technical skills;
- A comparative analysis of multiple large language models using clustering, substitution-coverage analysis, Mantel correlation, and Adjusted Rand Index;
- An interpretable representation of relationships between medical specialties, with potential applications in curriculum design and health system planning.

II. METHODS

A. Study Design

This study presents a computational analysis framework for constructing and analyzing specialty-to-specialty similarity matrices based on large language model (LLM) relevance ratings over multiple descriptor sets, followed by statistical and clustering analyses of the resulting similarity structures. The study is observational in nature and relies exclusively on publicly defined curricular and disease classification sources.

B. Setting

The study was conducted in 2025 through a collaboration between Hassan II University of Casablanca and the University of Franche-Comté. It focused on the 44 medical specialties officially recognized in France by the Arrêté du 21 avril 2017 [13] and incorporated disease information from the International Classification of Diseases, 11th Revision (ICD-11, 2025-01 release) [14].

C. Variables

The main variable was the pairwise semantic similarity between every pair of specialties for three independent criteria: diseases, educational teachings, and technical skills.

The framework is criterion-configurable: semantic similarity is defined with respect to the descriptor sets selected for analysis. In this study, diseases, educational teachings, and technical skills were used as three representative criteria, but

the choice of criteria is not fixed and may vary depending on data availability and the objectives of the implementing institution. This design allows the framework to capture different notions of proximity, such as overlap in treated diseases, shared technical competencies, or common training content. Consequently, the interpretation of similarity depends on the descriptor set used. A high disease-based similarity indicates overlap in disease-family relevance, whereas a high technical-skill similarity indicates overlap in procedural or practical competencies. The composite score should therefore be interpreted as an aggregated semantic proximity score, not as a direct measure of professional replaceability.

D. Data Sources and Measurement

Curricular information describing specialty competencies, teaching modules, and procedural acts was extracted from the French regulatory decree [13]. The extraction used Gemini 2.5 Pro for its large-context capacity, followed by manual verification to ensure completeness and accuracy. Disease-family data were obtained from the ICD-11 2025-01 dataset [14]. The curated lists of specialties, teachings, technical skills, and ICD-11 disease families used for analysis were compiled from these sources and used as input descriptors.

For each model, the relevance of every disease, teaching, and technical skill to each of the 44 specialties was evaluated through targeted queries. Relevance ratings were assigned on a 0–100 scale, where 0 indicated no relevance and 100 indicated a core condition, teaching item, or technical skill for the target specialty. Before similarity computation, all ratings were divided by 100 to obtain normalized scores in the [0,1] interval. Claude 4 Sonnet and GPT-4.1 were queried in batches of 18 descriptors using structured tool/function calls, while LLaMA 3.2 3B was queried item by item through a local Ollama interface using structured JSON output. All models were queried with temperature set to 0. The core prompt template was: "You are a clinical knowledge assistant. Given the medical specialty [specialty] and the descriptor(s) [descriptor list or item], assign a relevance score from 0 (not relevant) to 100 (core condition for that specialty). Return your answer only in the structured output format and do not add explanations."

The same specialty list, ICD-11 disease-family list, teaching-item list, and technical-skill list were used across all models. Ratings were validated structurally by checking that each model produced one score for every specialty–descriptor pair before similarity matrices were computed. The resulting normalized ratings were organized into matrices, where rows represent specialties and columns represent descriptor items. Similarity between two specialties was computed by comparing their rating patterns across all items, producing a score between 0 and 1 that reflects their semantic proximity. A separate 44 × 44 matrix was obtained for each criterion and each model, and the three criterion-specific matrices were averaged element-wise to yield a composite similarity matrix for every model. The model-generated rating outputs and derived similarity matrices (hereafter referred to as Dataset 2) were used for all statistical analyses.

For a given descriptor set (diseases, teaching items, or technical skills), pairwise similarity between two specialties A and B was computed from the normalized model-generated relevance ratings as one minus the mean squared difference across all descriptor items. Specifically, for a descriptor set of size n, similarity was defined as follows:

$$Similarity(A,B) = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(r_{A,i} - r_{B,i}\right)^2 \qquad (1)$$

where, $r_{A,i}$ and $r_{B,i}$ denote the normalized relevance ratings assigned by the model to item i for specialties A and B, respectively.

This formulation ensures bounded similarity values between 0 and 1 because all relevance ratings were normalized to the [0,1] interval before comparison. Higher values indicate greater semantic proximity. The squared-difference formulation was selected because it compares complete

relevance profiles item by item and penalizes large disagreements more strongly than small differences. The metric is simple, interpretable, and directly linked to the rating scale used in the study. Nevertheless, other similarity functions may emphasize different aspects of specialty proximity and should be examined in future sensitivity analyses.

Criterion-specific similarity matrices were then aggregated element-wise to produce a composite similarity matrix for each model.

The overall procedure is summarized in Algorithm 1.

Algorithm 1: Construction of specialty semantic similarity matrices using a large language model

Input:
- S = set of medical specialties
- $\{D_k\}$ = collection of descriptor sets, k = 1 … K
- M = large language model

Output
- $C_{composite}$ = composite similarity matrix

Initialize
- $\{R_k\}$ = collection of relevance ratings matrices, k = 1 … K
- $\{C_k\}$ = collection of per criterion similarity matrices, k = 1 … K

Compute
For each descriptor set D_k do
- For each specialty s in S do
  - For each descriptor item d in D_k do
    - Query M to obtain relevance rating $r_{s,d}$
    - Store rating $r_{s,d}$ in R_k
  - End
- End

End

For each ratings matrix R_k do
- For each pair of specialties (s_i, s_j) do
  - Compute $Similarity_k(s_i, s_j)$ using Eq. (1)
  - Store $Similarity_k(s_i, s_j)$ in C_k
- End

End

For each pair of specialties (s_i, s_j) do
- Compute:
- $C_{composite}[i][j] = (1 / K) * \sum_{k=1}^{K} C_k[i][j]$

End

E. Outcomes

Primary outcomes were 1) criterion-specific and composite 44×44 similarity matrices, 2) hierarchical clustering structures and cluster assignments, 3) substitution-coverage curves parameterized by similarity thresholds, and 4) disagreement visualizations between models. These outputs are intended as reusable components for curriculum mapping and planning-oriented analyses of specialty proximity and cross-coverage.

F. Bias

Prompts were standardized across models to reduce variation caused by prompt wording. Claude 4 Sonnet and GPT-4.1 were queried in batches of 18 descriptors using structured tool/function calls, whereas LLaMA 3.2 3B was queried item by item through a local Ollama interface because of its smaller model size and local execution setting. All models were queried with temperature set to 0.

The descriptor extraction workflow partly relied on Gemini 2.5 Pro for large-context document processing. To reduce extraction bias, the resulting descriptor lists were manually checked against the original French regulatory source for completeness and accuracy. However, no independent double-review process or inter-reviewer agreement metric was implemented.

Each descriptor item was queried once per model. Therefore, the analysis controls prompt format and temperature but does not estimate intra-model variability across repeated runs.

G. Study Size

The analysis included all 44 French medical specialties and their complete set of 843 descriptors: 307 disease families, 260 teaching items, and 276 technical skills. Because the study covers the entire population of specialties defined in the national framework, statistical sampling or inferential power estimation was not applicable.

H. Statistical Methods

All analyses were conducted in Python 3.11.12 (Python Software Foundation, Beaverton, OR, USA) using the open-source libraries NumPy 2.2.5, Pandas 2.2.3, SciPy 1.16.0, scikit-learn 1.7.0, and Matplotlib 3.10.1 (NumFOCUS, Austin, TX, USA).

Descriptive statistics were calculated for each model's similarity matrix, including the minimum, maximum, mean, and standard deviation of off-diagonal similarities. Hierarchical clustering with average linkage was applied to identify structural groupings of specialties. Agreement between models was evaluated using Mantel correlation tests with 999 permutations and the Adjusted Rand Index (ARI) to assess clustering consistency. Substitution-coverage analysis quantified, for each specialty, the proportion of potential substitutes above given similarity thresholds (0.85–1). The resulting matrices, dendrograms, and heatmaps were visualized to support quantitative and qualitative interpretation of inter-model consistency, specialty clustering, and substitution-coverage patterns.

III. RESULTS

This section presents the main findings of the proposed framework, including descriptive statistics, clustering structures, substitution-coverage patterns, and inter-model agreement analysis.

Overall, the results indicate that the proposed framework produces consistent and interpretable semantic representations of relationships between medical specialties across the evaluated large language models.

A. Descriptive Overview

All models produced complete 44 × 44 similarity matrices across the three criteria (diseases, teachings, and technical skills), as well as aggregated composite matrices. Table I presents the descriptive statistics of similarity coefficients for each model.

Similarity values were consistently high, with mean off-diagonal values around 0.87, indicating substantial semantic proximity across specialties. Because the rating distributions were dense, high absolute similarity values should be interpreted cautiously. The relative structure of the matrices, including clustering patterns and inter-model correlations, is more informative than the absolute magnitude of individual similarity scores. Claude 4 Sonnet and GPT-4.1 exhibited nearly identical statistical profiles, whereas LLaMA 3.2 3B produced slightly more homogeneous similarity distributions.

These results indicate strong agreement between Claude 4 Sonnet and GPT-4.1, with LLaMA 3.2 3B showing reduced variance and less differentiation between specialties.

TABLE I. DESCRIPTIVE STATISTICS OF SIMILARITY COEFFICIENTS PER MODEL

Model	Min	Max	Mean	SD
Claude 4 Sonnet	0.725	0.977	0.867	0.045
GPT-4.1	0.724	0.979	0.867	0.045
LLaMA 3.2 3B	0.761	0.967	0.867	0.038

B. Clustering of Specialties

Fig. 1 illustrates the resulting hierarchical clustering structure for Claude 4 Sonnet.

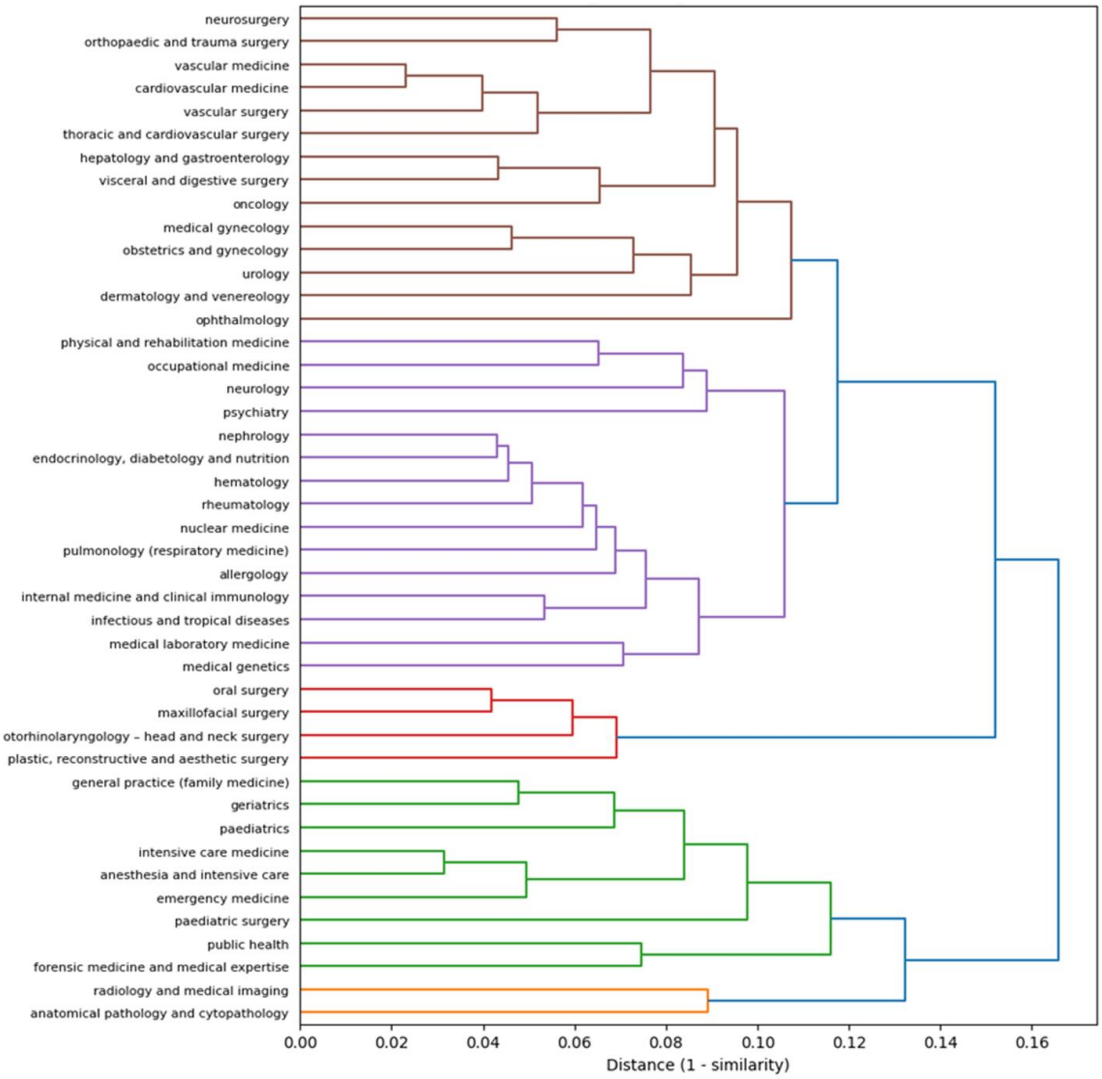


Fig. 1. Hierarchical clustering of 44 specialties (average linkage; distance = 1 – similarity) using the Claude 4 Sonnet composite matrix.

Hierarchical clustering applied to the Claude 4 Sonnet matrix revealed five major clusters of specialties corresponding to recognizable functional families within hospitals and universities. These groups include: 1) surgical and organ-based specialties; 2) internal medicine and diagnostic fields; 3) craniofacial surgical disciplines; 4) front-line, age-specific, and critical-care services; and 5) diagnostic imaging and pathology.

The identified clusters appear to correspond to coherent functional groupings observed in clinical practice and medical education. However, this interpretation remains qualitative and should be confirmed through structured expert validation.

C. Substitution-Coverage Analysis

Fig. 2 illustrates substitution coverage as a function of similarity threshold for Claude 4 Sonnet.

At a similarity threshold of 0.90, most specialties had between two and five semantically close candidates. This result does not imply automatic replacement between specialties, but it identifies pairs or groups where descriptor-based overlap may justify closer examination for short-term coverage, coordination, or referral-support scenarios.

As the threshold increased beyond 0.94, substitution options decreased sharply, highlighting areas where specialization limits potential interchangeability. These patterns suggest that semantic similarity can be used as an exploratory screening tool for identifying possible cross-specialty coverage candidates. Final assessment of substitutability would require additional clinical, regulatory, organizational, and patient-safety criteria beyond the semantic descriptors analyzed in this study.

D. Inter-Model Agreement

The Mantel test confirmed strong concordance between Claude 4 Sonnet and GPT-4.1 similarity matrices, with a correlation coefficient of r = 0.904 and p = 0.001. Correlations involving LLaMA 3.2 3B were lower, with Claude–LLaMA r = 0.683 and GPT–LLaMA r = 0.586.

Fig. 3 illustrates inter-model agreement across clustering resolutions.

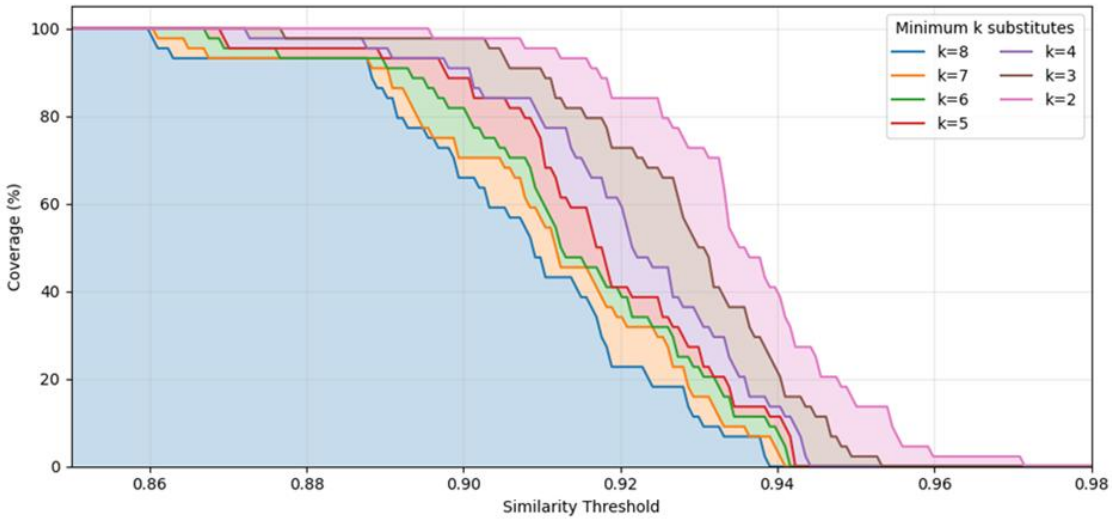


Fig. 2. Substitution coverage per similarity threshold (0.85 – 1.00) using the Claude 4 Sonnet composite matrix.

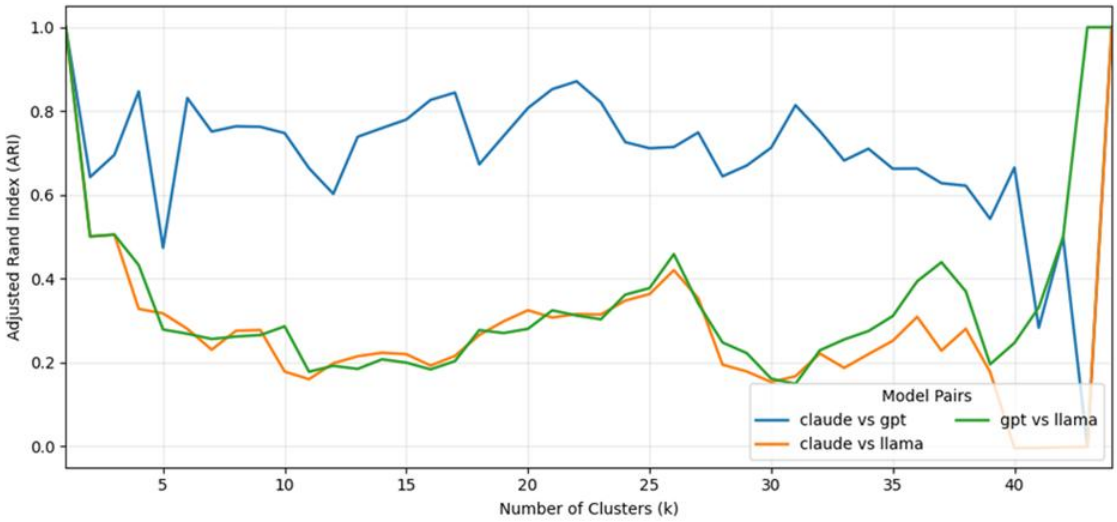


Fig. 3. Inter-model cluster agreement measured by Adjusted Rand Index (ARI) across k = 1-44.

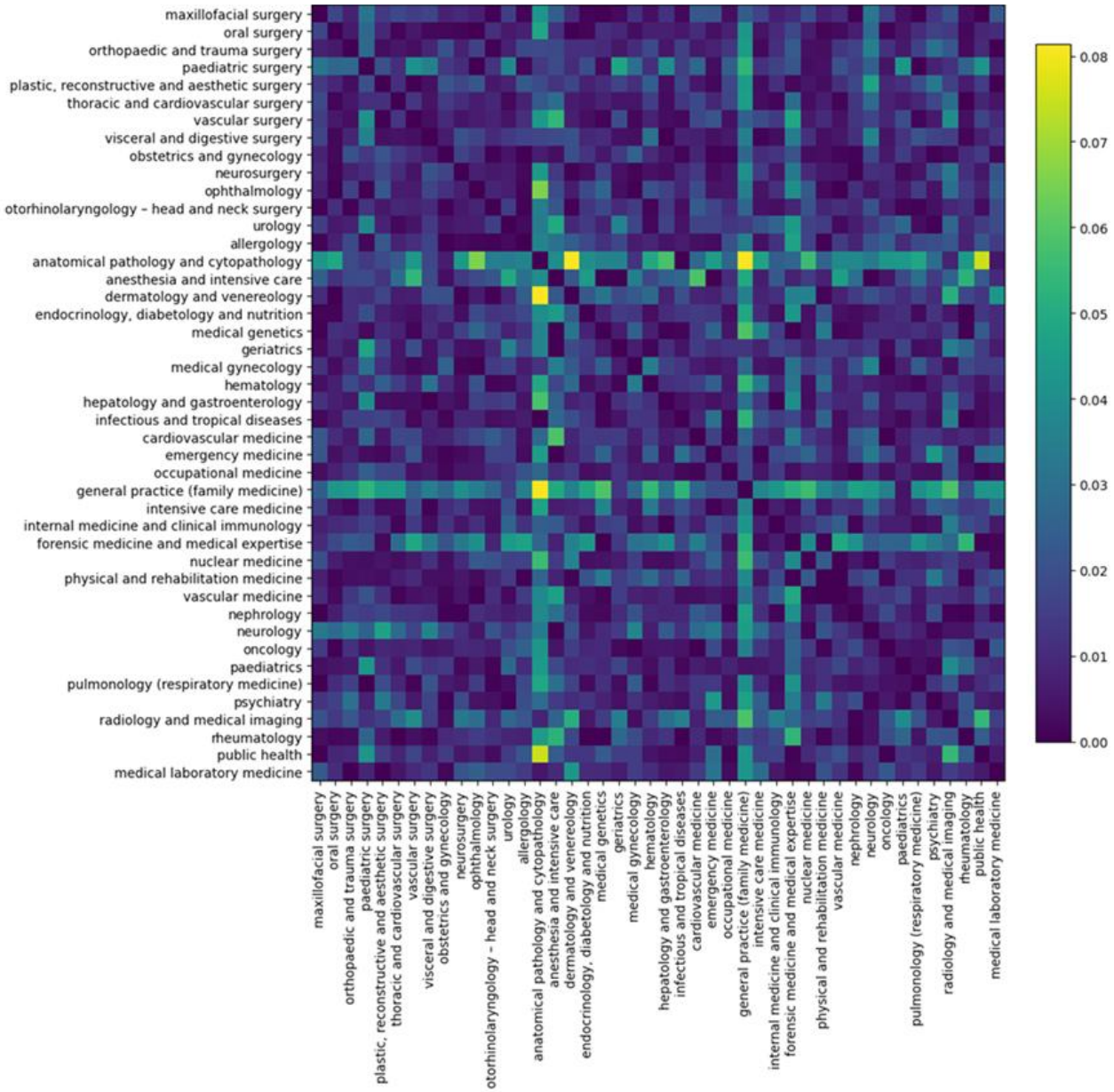


Fig. 4. Heatmap of absolute differences between Claude 4 Sonnet and GPT-4.1 composite similarity matrices (0-1 scale).

Cluster agreement, measured using the Adjusted Rand Index across all possible numbers of clusters (k = 1–44), remained consistently high between Claude 4 Sonnet and GPT-4.1, while LLaMA 3.2 3B diverged more at finer resolutions.

This confirms that large-scale proprietary models capture highly consistent semantic structures, while smaller models preserve global patterns with reduced granularity.

E. Model Disagreements

Fig. 4 illustrates the absolute differences between composite similarity matrices of Claude 4 Sonnet and GPT-4.1.

A heatmap comparing the two models revealed localized areas of divergence, particularly for general practice and anatomical pathology. These discrepancies correspond to specialties whose definitions span multiple domains.

These localized differences highlight areas where semantic boundaries between specialties are less clearly defined, indicating potential zones requiring expert validation.

IV. Discussion

A. Key Results

This study suggests that large language models can be used to construct consistent and interpretable computational representations of semantic relationships between medical specialties. All evaluated models produced coherent similarity structures, with strong agreement observed between Claude 4 Sonnet and GPT-4.1 across statistical measures. The resulting clustering revealed groupings of specialties that appear consistent with functional domains in clinical practice and medical education, although independent expert validation remains necessary.

B. Interpretation

The high level of agreement between Claude 4 Sonnet and GPT-4.1 suggests that large-scale instruction-tuned models capture consistent semantic structures when applied to

structured domain descriptors. In contrast, LLaMA 3.2 3B generated smoother similarity distributions, indicating reduced differentiation while preserving overall structural coherence. These differences likely reflect variations in model scale, architecture, and training data exposure.

From a computational perspective, the results indicate that semantic similarity derived from LLM-based ratings can provide a useful exploratory representation of structural relationships between domain entities. The multi-criteria aggregation approach further enhances interpretability by combining complementary perspectives on specialty similarity.

C. Comparison with Previous Work

Previous research on large language models in healthcare has primarily focused on clinical applications, diagnostic use cases, and evaluation tasks such as medical examination performance [2], [3], [7]. In parallel, several studies have explored their implications for medical education and curriculum design [1], [6], [9].

Other approaches have applied artificial intelligence to the structuring of medical knowledge, including specialty or subdomain classification [10], [11], as well as the analysis of semantic relatedness between biomedical concepts [12].

However, few studies have addressed the problem of modeling relationships between medical specialties as a structured computational task. In contrast, this work focuses on relational modeling between specialties, providing a quantitative and interpretable representation of their semantic proximity.

D. Limitations

First, as with all studies involving large language models, the results represent a snapshot of model behavior during a specific period. Subsequent model updates, retraining, or changes in deployment settings may alter response patterns and similarity values.

Second, each descriptor item was queried once per model without repeated sampling, so intra-model variability was not formally assessed. Although the temperature was set to 0, repeated runs could help quantify the stability of rating outputs.

Third, high absolute similarity values should be interpreted in relation to the rating scale and descriptor distribution. Because the model-generated ratings formed relatively dense profiles, the relative structure of the matrices, clustering patterns, and inter-model correlations are more informative than isolated similarity magnitudes. Future work should include randomized baselines, perturbation-based sensitivity analysis, and bootstrap confidence intervals to further assess robustness.

Fourth, the study focused exclusively on the 44 medical specialties defined in the French regulatory framework, which may limit direct applicability to other national healthcare systems, specialty taxonomies, or subspecialty structures.

Fifth, the interpretation of clusters was qualitative and was not formally validated by an independent clinical or educational expert panel. Future work should compare the computational clusters with structured expert assessments of specialty proximity.

Finally, substitution coverage was analyzed as an exploratory semantic indicator. Actual short-term coverage or substitution decisions require additional criteria, including clinical responsibility, regulatory authorization, institutional protocols, available competencies, and patient-safety considerations.

E. Generalizability

The proposed framework is designed to be adaptable to different contexts and data availability. The choice of descriptor criteria is not fixed and may vary depending on the objectives and constraints of the implementing entity. In this study, diseases, educational teachings, and technical skills were selected as representative criteria; however, alternative or additional descriptors could be incorporated depending on the use case.

This flexibility allows the framework to address semantic similarity in a context-dependent manner, effectively answering the question “similar with respect to what?”. For instance, using disease-based descriptors highlights specialties managing similar clinical conditions, technical skill descriptors capture overlap in procedural competencies, and educational descriptors reflect shared training structures.

As a result, the framework can be customized to emphasize different aspects of specialty relationships, making it applicable across healthcare systems, educational institutions, and planning scenarios with varying priorities.

However, adaptation to another setting requires reconstruction of the descriptor sets using local specialty definitions, curricula, clinical responsibilities, and regulatory constraints. The framework should therefore be viewed as a configurable semantic analysis tool rather than a universal model of medical workforce substitutability.

V. Conclusion

This study presented a computational framework for modeling semantic relationships between medical specialties using large language models. By combining multiple descriptor sets, the proposed approach produces interpretable similarity matrices that capture structural relationships between specialties.

The results showed strong agreement between large-scale models and suggested that semantic similarity derived from LLM-based relevance profiles can provide meaningful exploratory representations of specialty proximity. These findings highlight the potential of large language models as tools for structured analysis in complex medical and educational domains.

The flexibility of the framework allows adaptation to different contexts and objectives, making it potentially useful for exploratory analyses in medical education, cross-specialty coordination, and preliminary healthcare planning. Future work

should incorporate repeated model sampling, robustness analysis, randomized baselines, and structured expert validation to further assess the stability and practical relevance of the proposed similarity structures.

Declaration on Generative AI

The authors acknowledge the use of generative AI tools to assist with language refinement, wording revision, and manuscript editing during the preparation of this article. Their use was limited to writing assistance and did not determine the scientific ideas, study design, methodology, analyses, interpretation of results, or conclusions. All AI-assisted text was critically reviewed, edited, and validated by the authors, who take full responsibility for the originality, accuracy, and integrity of the manuscript.

References

[1] X. Xu, Y. Chen, and J. Miao, "Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review," Journal of Educational Evaluation for Health Professions, vol. 21, Art. no. 6, 2024, doi: 10.3352/jeehp.2024.21.6.

[2] E. N. Liang, S. Pei, P. Staibano, and B. van der Woerd, "Clinical applications of large language models in medicine and surgery: a scoping review," Journal of International Medical Research, vol. 53, no. 7, Art. no. 03000605251347556, 2025, doi: 10.1177/03000605251347556.

[3] H. Su, Y. Sun, R. Li, A. Zhang, Y. Yang, F. Xiao, Z. Duan, J. Chen, Q. Hu, T. Yang, B. Xu, Q. Zhang, J. Zhao, Y. Li, and H. Li, "Large language models in medical diagnostics: scoping review with bibliometric analysis," Journal of Medical Internet Research, vol. 27, Art. no. e72062, 2025, doi: 10.2196/72062.

[4] A. M. Alkalbani, A. S. Alrawahi, A. Salah, V. Haghighi, Y. Zhang, S. Alkindi, and Q. Z. Sheng, "A systematic review of large language models in medical specialties: applications, challenges and future directions," Information, vol. 16, no. 6, Art. no. 489, 2025, doi: 10.3390/info16060489.

[5] X. Chen, J. Xiang, S. Lu, Y. Liu, M. He, and D. Shi, "Evaluating large language models and agents in healthcare: key challenges in clinical applications," Intelligent Medicine, vol. 5, no. 2, pp. 151–163, 2025, doi: 10.1016/j.imed.2025.03.002.

[6] H. C. Lucas, J. S. Upperman, and J. R. Robinson, "A systematic review of large language models and their implications in medical education," Medical Education, vol. 58, no. 11, pp. 1276–1285, 2024, doi: 10.1111/medu.15402.

[7] B. C. Torres-Zegarra, W. Rios-Garcia, A. M. Ñaña-Cordova, K. F. Arteaga-Cisneros, X. C. B. Chalco, M. A. B. Ordoñez, C. J. G. Rios, C. A. R. Godoy, K. L. T. P. Quezada, J. D. Gutierrez-Arratia, and J. A. Flores-Cohaila, "Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study," Journal of Educational Evaluation for Health Professions, vol. 20, Art. no. 30, 2023, doi: 10.3352/jeehp.2023.20.30.

[8] B. C. Gin, O. ten Cate, P. S. O'Sullivan, K. E. Hauer, and C. Boscardin, "Exploring how feedback reflects entrustment decisions using artificial intelligence," Medical Education, vol. 56, no. 3, pp. 303–311, 2022, doi: 10.1111/medu.14696.

[9] O. Ng, Z. H. Tay, L. V. E. Wilding, K. B. Ng, and S. P. Han, "Transforming curriculum mapping: a human-AI hybrid approach," Medical Education, vol. 58, no. 5, pp. 582–583, 2024, doi: 10.1111/medu.15331.

[10] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," BMC Medical Informatics and Decision Making, vol. 17, no. 1, Art. no. 155, 2017, doi: 10.1186/s12911-017-0556-8.

[11] C. Mao, Q. Zhu, R. Chen, and W. Su, "Automatic medical specialty classification based on patients' description of their symptoms," BMC Medical Informatics and Decision Making, vol. 23, no. 1, Art. no. 15, 2023, doi: 10.1186/s12911-023-02105-7.

[12] Y. Mao and K. W. Fung, "Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts," Journal of the American Medical Informatics Association, vol. 27, no. 10, pp. 1538–1546, 2020, doi: 10.1093/jamia/ocaa136.

[13] République Française, "Arrêté du 21 avril 2017 relatif aux connaissances, aux compétences et aux maquettes de formation des diplômes d'études spécialisées et fixant la liste de ces diplômes et des options et formations spécialisées transversales du troisième cycle des études de médecine," Ministère de l'Enseignement supérieur et de la Recherche, Paris, France, 2017. [Online]. Available: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034502881

[14] World Health Organization, "International Classification of Diseases, 11th Revision (ICD-11)," Geneva, Switzerland, 2025. [Online]. Available: https://icd.who.int/

An Intelligent Semantic-Aware Academic Library System Using Natural Language Processing and Knowledge Graphs

Mohamad Shady Alrahhal[1], Bandar Abdulwahed Fallatah[2], Omar F. Aloufi[3], Abdullah M. Barakeh[4]
AIMD Laboratory-Computing College, Fahad Bin Sultan University, Tabuk City, KSA[1, 4]
The Boeing Company, Tabuk City, KSA[2]
Information Systems Department-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia[3]
Department of Information Systems-College of Computer Science and Engineering,
Taibah University, Madinah, Saudi Arabia[3]

***Abstract*—Traditional library systems designed for academic research suffer from poor integration when employing AI-based techniques for enhancement purposes. These limitations are mainly related to ignoring semantic relationships during search, overlooking hidden relationships among words, and lacking the generalization capability required to enhance researchers' experience. To address these issues, the Intelligent Library System (IntLS) is proposed and further enhanced through effective knowledge graph modeling to enable more accurate retrieval of results. In addition, the basic NLP processing steps are optimized to capture hidden relationships during tokenization, stop-word removal, and stemming stages. The architecture of the proposed system consists of two main components: a semantic component, responsible for generating semantic representations of documents, and an analytics component, responsible for analyzing historical searches to predict future user needs and support effective resource management. The proposed system is compared with two well-known systems using similarity-based accuracy and a set of AI-based evaluation metrics. The enhanced system, Enh-IntLS, demonstrates superior performance, achieving a 1.9% improvement in similarity-based accuracy and a 1.7% improvement in AI-based accuracy.**

***Keywords*—Library systems; academic search; generalization; knowledge graph; similarity; accuracy**

I. Introduction

Traditional Library Systems (TLSs) are essentially digital catalogues and circulation management platforms that provide limited analytics. They rely mainly on reactive usage statistics rather than predictive capabilities. In other words, traditional library management systems are database-driven systems that primarily digitize and automate manual processes. Such automation benefits stakeholders (i.e., both librarians and users) by enabling them to manage book records, track borrowing and returning transactions, monitor due dates, and generate basic reports [1]. In light of the rapid advancement of technologies, library system developers have begun employing artificial intelligence to enhance the quality of services provided to users. However, the integration of artificial intelligence also opens the door to new challenges.

A. Motivation

Despite these improvements, traditional library management systems still exhibit several limitations within modern information environments [2, 3]. First, their search mechanisms are primarily keyword-based, lacking semantic understanding and intelligent recommendation capabilities. As a result, users often struggle to find relevant resources beyond exact keyword matches. Second, these systems typically operate reactively rather than proactively; they do not predict user needs, recommend resources based on borrowing history, or dynamically optimize inventory. Third, traditional LMS platforms provide limited data analytics, which restricts libraries from extracting actionable insights related to user behavior, demand forecasting, and collection development. Furthermore, they lack automation in tasks such as intelligent cataloging, classification, anomaly detection (e.g., identifying overdue patterns), and user personalization. In many cases, decision-making remains dependent on manual librarian intervention, which reduces scalability and efficiency in large academic institutions.

B. Statement of Problem

With the advancement of Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and data analytics technologies, new opportunities have emerged to transform library services into intelligent, adaptive, and user-centric systems. AI-powered systems can enable semantic search, automated metadata generation, personalized recommendation engines, and predictive analytics for resource demand [4]. These capabilities can significantly enhance the user experience, improve operational efficiency, and support strategic resource management.

In other words, there remains a gap between conventional rule-based LMS platforms and fully intelligent, AI-driven library ecosystems. Many existing systems do not integrate advanced AI components within a unified and scalable architecture capable of supporting intelligent search, predictive decision-making, and adaptive service personalization. Bridging this gap is critical for establishing smart library systems that enable researchers to accurately find the documents they need. Moreover, to the best of our knowledge,

the evaluation of the generalization quality attribute has not been addressed in previous works, which is important for enhancing researchers' trust in academic search systems.

C. Research Questions

Based on the problem stated above, the following research questions are formulated:

- How can intelligent search be enabled based on the semantics of keywords?
- How can predictive analytics be enabled for resource demand?

D. Contributions

This work provides the following contributions:

- Proposes the Intelligent Library System (IntLS) to address the research questions mentioned above. The system employs an NLP-based approach in its development.
- Enhances IntLS by developing a knowledge graph of library topics, enabling accurate predictive analytics rather than relying solely on descriptive analytics.
- Evaluating the generalization quality attribute to demonstrate the effective applicability of the proposed system in real-world environments.
- Conducts a comparative evaluation with well-known smart library systems to demonstrate the superiority of the proposed approach.

E. Work Organization

The rest of this study is organized as follows: Section II presents an overview of both Traditional Library Systems (TLSs) and Intelligent Library Systems (ILSs). Section III describes the proposed system in detail. The experimental results are presented and discussed in Section IV. Finally, Section V concludes the study.

II. Related Work

This section presents a brief overview of both traditional library systems and smart library systems.

A. Traditional Library Systems

Authors of work [5] stated that the need for a specialist academic library system is pressing to enhance the academic productivity of researchers. The previous work is supported by another article provided in [6], where a comparison between traditional and smart library systems is provided. The final result of this work stated that artificial intelligence techniques can play a significant role in enhancing the accuracy of academic search. The Ex Libris Alma platform [7], widely adopted by academic and research libraries worldwide, represents a recent generation of library services platforms that integrate cataloging, fulfillment, and resource management into a unified cloud-based system. According to the official Ex Libris documentation, Alma incorporates data analytics capabilities to improve operational efficiency in resource management. However, its analytics remain primarily descriptive rather than predictive, and the platform relies on proprietary infrastructure with associated licensing costs that may not be accessible to all academic institutions, particularly those in developing regions.

B. Intelligent Library Systems

Many techniques are employed to develop smart library systems, such as text classification and natural language processing. In the work [8], the authors proposed a text classification-based system supported by the KNN algorithm. The key idea is to create clusters of text according to categories generated based on the smart recommendation of text classification. authors of work [9] presented an NLP-based library system supported by a correlation function to handle semantic issues. The key idea is to employ correlation based on the domain where the text belongs. The correlation function provides statistical information about semantics after following standard NLP steps. However, the limitation of this work is related to ignoring the knowledge graph database.

It is worth mentioning that KNN-based and NLP-based systems are involved in the comparison with the proposed system of this study.

III. Proposed System

This section presents the proposed intelligent system, referred to as the Intelligent Library System (IntLS), as well as the enhanced one. First, the architecture of the system is introduced. Then, the role of each component within the architecture is described in detail. The testing of the system and the corresponding discussion of results are presented in a separate section.

A. IntLS Architecture

The architecture of IntLS is composed of two main components: the semantic component and the predictive analytics component. Fig. 1 below illustrates the role of the first component, in which several techniques are employed to generate the desired output.

To enable semantic-based search, the semantic component performs several steps, as described below.

1) Semantic representation of documents: The output of this step is to capture the meaning of documents rather than relying solely on exact word matches. This enables users to retrieve results based on context-aware queries. For example, if a user searches for books on artificial intelligence, resources containing topics such as deep learning, machine learning, and text mining will also appear in the result set. To achieve this capability, library resources (e.g., books, articles, theses, etc.) must first be transformed into semantic representations. To perform this task, text preprocessing is applied to important textual fields of the documents, such as the title, abstract, and keywords. Text preprocessing includes the following sub-tasks:

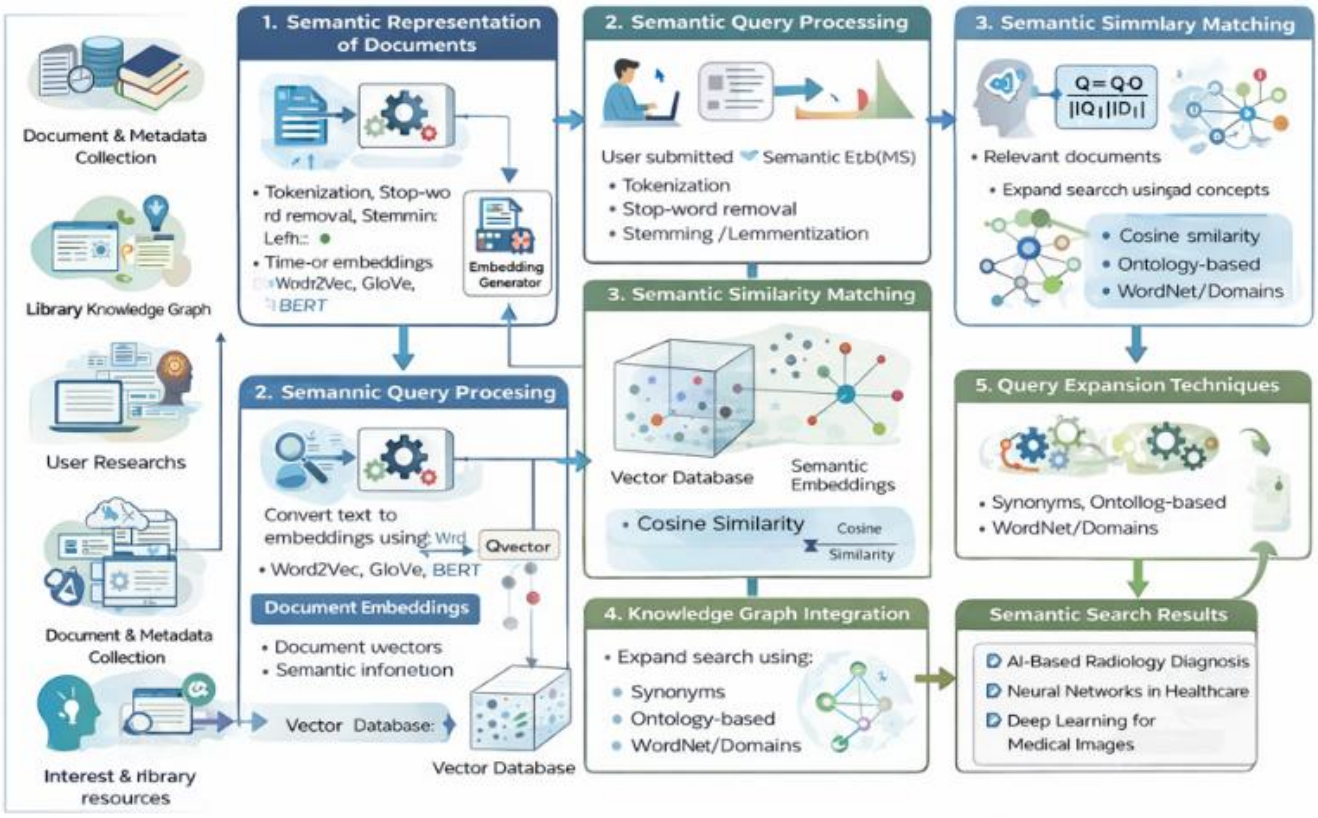


Fig. 1. Role of semantic component.

- Tokenization: This step involves converting a sentence into a sequence of words, commonly referred to as "tokens." Tokens represent the basic building blocks upon which text analysis and other processing methods are performed. Using the title of work [10], the output will be as follows:

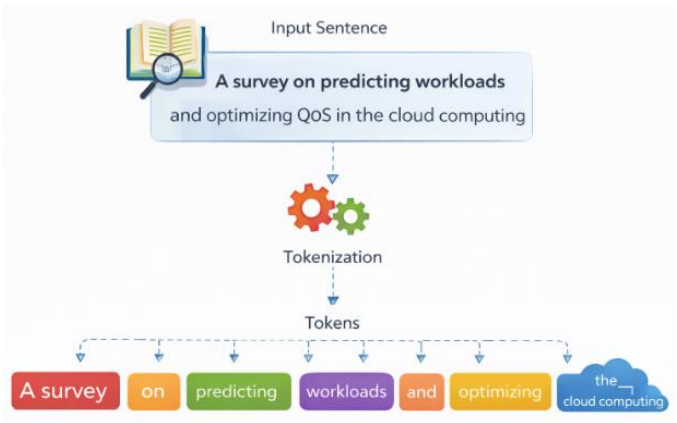


Fig. 2. Output of the tokenization step.

As shown in Fig. 2, a whitespace-based method is used to generate tokens. However, this work enhances this method by considering strong relationships between words rather than treating them individually. Therefore, the first token includes "A survey," and the last token includes "cloud computing," as they represent integral phrases within the computer science domain. In addition, abbreviations such as "QoS" in our example are removed because they may introduce ambiguity and negatively affect accuracy.

- Stop-word removal: Researchers have provided a list of such words in the Bag-of-Words (BoW) library. The objective of this step is to eliminate words that have low importance in conveying meaning (i.e., words that have minimal impact on the overall semantics). Fig. 3 illustrates the output of the same previous example.

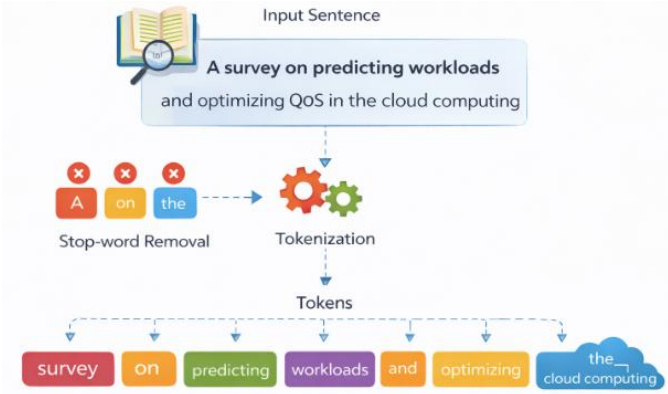


Fig. 3. Output of stop-word removal step.

As shown in Fig. 3, the words "A," "on," and "the" are removed. However, this work introduces an enhancement that considers the relationship conveyed by the word "and." This is because authors of a document often use "and" to emphasize the relationship between two concepts (e.g., "workloads" and

"optimizing"), highlighting a form of trade-off or association between the two terms.

- Stemming: This is the process of converting words into their base form, or stem. In this study, Porter's algorithm is used for this purpose. The output of this step is illustrated in Fig. 4.

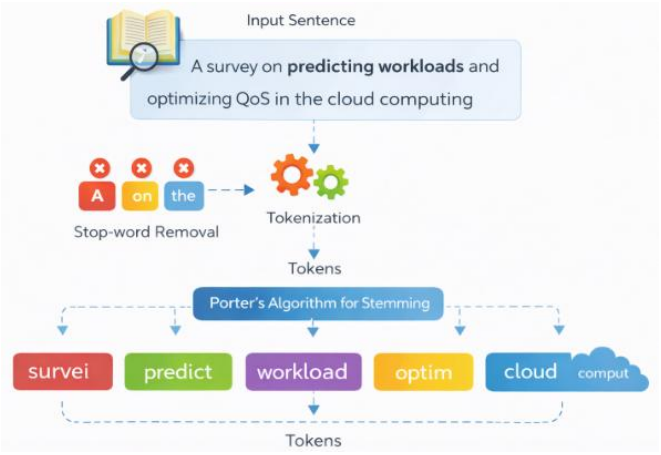


Fig. 4. Output of the stemming step.

- Convert text into semantic vectors (embeddings): This step uses the Word2Vec model to transform textual data into semantic vector representations, as shown in Fig. 5. The objective of this step is to map the text into a numerical vector space, enabling similarity calculations in subsequent stages.

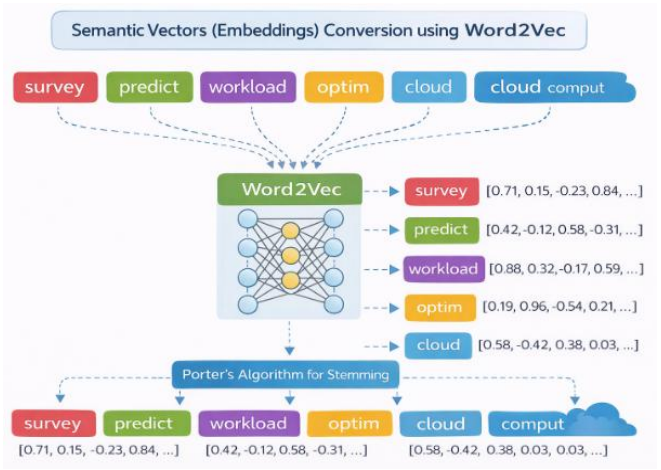


Fig. 5. Output of embedding step.

2) Semantic query processing: The objective of this step is to enable IntLS to retrieve documents with similar meanings even when different keywords are used. For example, if a user submits the query "intelligent learning for the medical field," the system can retrieve studies such as "deep learning in medical imaging" or "neural network–based COVID-19 diagnosis."

In this context, the same preprocessing steps used in the previous example are applied (i.e., tokenization, stop-word removal, stemming, etc.). The query is then converted into a semantic embedding vector. Finally, semantic similarity matching is performed, as described below.

3) Semantic similarity matching: The objective of this step is to rank the retrieved documents within the result set based on their semantic closeness to the query. In this context, the Cosine Product similarity measure [11] is used. It is defined as follows:

$$Sim\ (D,Q) = \frac{D.Q}{\|D\| \times \|Q\|} \quad (1)$$

where, Q is the query vector, and D is the document vector.

Problem. A critical issue arises when considering the intersections among different domains. For example, the computer vision domain is tightly coupled with deep learning, while artificial intelligence is strongly linked with text and data mining. In such scenarios, the accuracy of retrieval results may decrease due to overlapping or related concepts across domains. To address this challenge, this work introduces an enhancement to IntLS through the integration of a Knowledge Graph [12].

B. Enhanced IntLS Architecture (Enh-IntLS)

The key idea behind this enhancement is the construction of a knowledge graph of library topics. The objective of this enhancement is to support semantic search, recommendation, topic discovery, and trend analysis across intersecting domains, which is essential for researchers seeking relevant documents. A knowledge graph is a network of nodes and edges, where nodes represent entities such as authors, topics, journals, and departments, while edges represent relationships such as *written by*, *belongs to*, *related to*, and *published in*. Through this structure, Enh-IntLS can understand how topics, books, authors, users, and research areas are interconnected. Fig. 6 illustrates the basic steps involved in creating the knowledge graph.



Fig. 6. Steps of knowledge graph development.

From mathematical modeling, the basic steps are provided as follows:

Let the final knowledge graph be: $G = (V, E)$ (2)

where, V is the set of entities (nodes), and E is the set of relationships (edges). The scope S ∈ D, where D is all possible domains, such as cloud computing.

The graph schema includes the following entity types:

V={Publication, Author, Topic, FieldOfStudy, Journal, Conference, Department, Institution}

The main relation types are defined as follows:

E={written_by,published_in,belongs_to,has_topic,cites,affliated_with,related_to}

Important entities extracted from documents are represented by:

$$E = \{e_1, e_2, e_3, .., e_n\} \quad (3)$$

Entities are extracted from structured and semi-structured academic metadata. The main input fields include publication titles, abstracts, keywords, authors, publication venues, departments, institutions, citation links, and fields of study. The extraction process follows these steps: First, the textual fields of each document are preprocessed using tokenization, stop-word removal, phrase detection, and stemming. Second, candidate entities are identified from titles, abstracts, and keywords. Third, structured metadata fields, such as author names, journal names, conference names, institutional affiliations, and fields of study, are directly extracted as graph entities.

After extraction, each entity is linked to an existing graph node or inserted as a new node if no match is found. Entity linking is performed using three steps. First, entity labels are normalized by converting them to lowercase, removing punctuation, and standardizing abbreviations. Second, exact matching is used to link identical entity names. Third, semantic matching is applied using embedding similarity to link conceptually equivalent or highly related entities.

Between two different entities, the relationship is defined as:

$$E = \{(e_i, r, e_j)\} \quad (4)$$

where, e_i, e_j are entities, and r is kind of relationship.

To collect data from different resources, let:

$$X = \{x_1, x_2, x_3, \ldots x_k\} \quad (5)$$

Refers to the set of resources. Then, each document x can be represented by:

$$x = \{title, abstract, keywords\} \quad (6)$$

E_i is the extracted entity from a document x. Then, the relationship extracted from the pairs is:

$$r = relation\ (e_i, e_j) \quad (7)$$

Relations are extracted using both metadata-based and semantic-based methods. Metadata-based relations are obtained directly from structured fields. For example, if publication (P_i) has author (a_i) the triple $(P_i, written_by, a_i)$ is created.

The relationships are embedded within the graph. However, the graph is stored using three parameters as follows:

$$StrG = (s, p, o) \quad (8)$$

where, S is the subject, P is the predicate, and O is the object.

For example, (Article A covers Cloud Computing). Based on this representation, queries are issued from the side of the user, so that searching relevant nodes for a query is modeled as:

$$Result = argmax_{e \in E} Score\ (q, e) \quad (9)$$

To make the enhancement clearer, the following provides an example of query enrichment: Assume that a user submits the query: “deep learning in medical imaging”. First, the query is preprocessed and converted into a semantic vector. The embedding component retrieves candidate documents that contain semantically close expressions such as “neural networks for diagnosis,” “CNN-based image classification,” and “AI-based radiology analysis.” Then, the query entities, deep learning and medical imaging, are linked to corresponding nodes in the knowledge graph. Through graph traversal, Enh-IntLS identifies related nodes such as neural networks, image classification, medical diagnosis, radiology, and computer vision. Candidate documents connected to these related nodes receive higher graph-based relevance scores. For example, a document titled “Neural Network-Based COVID-19 Diagnosis from Chest Images” may be ranked higher even if it does not contain the exact phrase “deep learning in medical imaging.”

After the enhancement based on the knowledge graph, the predictive analytics component begins its function. The main objective of this component is to analyze historical usage data and semantic information about library resources to forecast future trends and user needs. To accomplish this task, the predictive analytics component performs the steps described below.

1) Predicting user information needs: By analyzing previous searches and browsing history, a prediction profile is established to identify resources that users may need in the future.

2) Demand forecasting for library resources: This step predicts which books, journals, or digital resources will be highly in demand in the future.

Mathematically, these actions are modeled as follows:

Let UP is user profile features, RF is resource features (topics, keywords, metadata), HID is historical interaction data, and f is predictive model. Then, predicted demand (PD) is:

$$PD = f(UP, RF, HID) \quad (10)$$

Fig. 7 illustrates the integration of the two components to generate the final output.

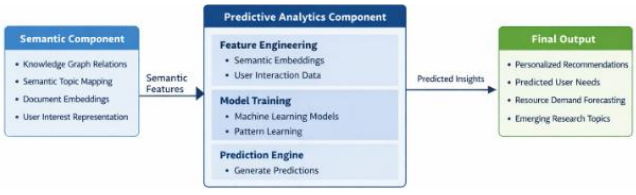


Fig. 7. Integration of the semantic component and predictive analytics component.

IV. Experimental Results and Evaluation

To evaluate the proposed IntLS and Enh-IntLS systems, the most recent Microsoft Academic Graph (MAG) dataset [13] was selected for training and testing. The dataset contains 109,105 triples. It is a heterogeneous graph that includes scientific publication records, citation relationships among these publications, as well as information about authors, institutions, journals, conferences, and fields of study. Therefore, it is considered a suitable benchmark dataset for conducting experiments. The dataset was divided into 80% for training and 20% for testing.

Evaluation is arranged to cover two main tasks, which are: 1) Semantic retrieval of academic resources; and 2) Resource demand prediction. For each task, input data, target output, ground truth labels, and evaluation metrics are defined as follows:

1) For the semantic retrieval task, the input consists of academic resource metadata, including titles, abstracts, keywords, topics, and knowledge-graph triples extracted from the Microsoft Academic Graph dataset. The target output is a ranked list of academic resources relevant to a submitted query. The relevance criterion is based on semantic overlap between the query concepts and the document concepts represented through Word2Vec embeddings and knowledge-graph relationships. The evaluation uses similarity-based accuracy to measure whether the retrieved resources are semantically aligned with the expected relevant topics.

2) For the demand prediction task, the input consists of simulated historical interaction records derived from academic-resource usage indicators, including resource-topic frequency, query-topic occurrence, and resource-demand class labels. The target output is a binary classification of future demand, namely correct/high demand and incorrect/low demand. The ground truth labels are represented through demand classes used in the confusion matrix. The evaluation uses accuracy, precision, recall, specificity, and ROC-based analysis.

It is worth mentioning that since the Microsoft Academic Graph is primarily a scholarly metadata and citation graph, it is integrated with simulated demand data to cover the evaluation of the second task. In other words, simulated interaction and synthetic demand-label data are used for predictive analytics. Specifically, simulated historical interaction records were constructed from MAG-based academic-resource indicators. These indicators included topic frequency, resource-topic occurrence, citation activity, field-of-study frequency, and graph-based relationships among publications, authors, and research areas.

This integration leads to update number of records in the selected dataset to be 214,000 records. In other words, the value of 214,000 evaluated instances refers to simulated demand-observation records generated from resource-topic occurrences, citation activity, field-of-study frequency, and synthetic interaction indicators. These simulated records were used only for evaluating the demand-prediction task, not for evaluating the semantic retrieval task.

B. Strategy of Evaluation and Used Metrics

We first evaluate the main system (i.e., IntLS) in comparison with the enhanced system (i.e., Enh-IntLS). Then, Enh-IntLS is compared with KNN-based and NLP-based systems reviewed in the related work section. In addition, the generalization quality attribute is evaluated by varying the training-to-testing ratios. It is worth mentioning that in this work, generalization is considered internal validation, as no real data is available. In future work, we tend to prove generalization using real data. For forecasting resource demand, the logistic regression algorithm [14] is employed.

Regarding the evaluation metrics, similarity-based accuracy is used to assess semantic similarity matching based on Eq. (1). Specifically, we define the semantic retrieval task as a top-k ranking problem. For each query, the system retrieves and ranks academic documents according to the semantic similarity score between the query vector and the document vector. In IntLS, this score is computed using Word2Vec-based document/query embeddings and Cosine Product similarity. In Enh-IntLS, the similarity score is further enriched using knowledge-graph relationships among topics, authors, journals, departments, and publications. A retrieved document is considered relevant if its semantic similarity score is greater than or equal to a predefined threshold and if it shares at least one expected topic, field-of-study, or knowledge-graph relationship with the query. In the revised manuscript, we define the threshold value as **0.70** after normalizing similarity scores into the range [0, 1]. Therefore, a retrieved document is considered correct when it satisfies the relevance criterion, and it is considered incorrect when it does not satisfy this criterion.

In addition, accuracy, precision, recall, and specificity are used to evaluate the prediction of resource demand. These metrics are derived from the confusion matrix [15] summarized in Table I.

TABLE I. Confusion Matrix

Predicted Dem Actual Dem	PD_{final}	$\neg PD_{final}$	Sum
PD_{final}	TP	FN	S
$\neg PD_{final}$	FP	TN	Su
Sum	$\widetilde{S}$	$\widetilde{Su}$	All

where:

- True positives (TPs): The positive demands that were correctly recommended by the system.
- True negatives (TNs): The negative demands that were correctly recommended by the system.
- False positives (FPs): The negative demands that were incorrectly recommended as positives by the system.
- False negatives (FNs): The positive demands that were recommended as negatives by the system.

$$\text{Accuracy} = \frac{TP+TN}{All} \quad (11)$$

$$\text{Precision} = \frac{TP}{\widetilde{S}} \quad (12)$$

$$\text{Recall} = \frac{\text{TP}}{\text{S}} \quad (13)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{Su}} \quad (14)$$

C. Similarity-Based Accuracy Evaluation of IntLS and Enh-IntLS (Semantic Retrieval Task)

In this subsection, the semantic retrieval task is evaluated, and the corresponding results are discussed. Table II presents the results obtained under varying training-to-testing ratios for evaluating the generalization quality attribute.

TABLE II. SIMILARITY-BASED ACCURACY RESULTS

System	Training to Testing Ratio	Sim Metric	AVG
IntLS	[80 to 20]	80.2	≈ 78.58
	[70 to 30]	80.1	
	[60 to 40]	79	
	[50 to 50]	75	
Enh-IntLS	[80 to 20]	93.3	≈ 88.83
	[70 to 30]	91	
	[60 to 40]	88	
	[50 to 50]	83	

Discussion. When the training ratio was decreased by 10%, and the testing ratio was increased by 10%, both systems showed decreasing values in the similarity-based accuracy metric. This can be explained by the fact that incorporating more records into the training dataset leads to the extraction of more features, which can be utilized during the similarity matching process, thereby resulting in higher accuracy values. However, Enh-IntLS demonstrated better performance compared to IntLS. This improvement can be attributed to the use of a knowledge graph of topics, which models similarity based on the semantics of input queries and retrieved results. Regarding the generalization quality attribute, Fig. 8 illustrates the difference between the baseline training-to-testing ratio and the average results obtained for both systems.

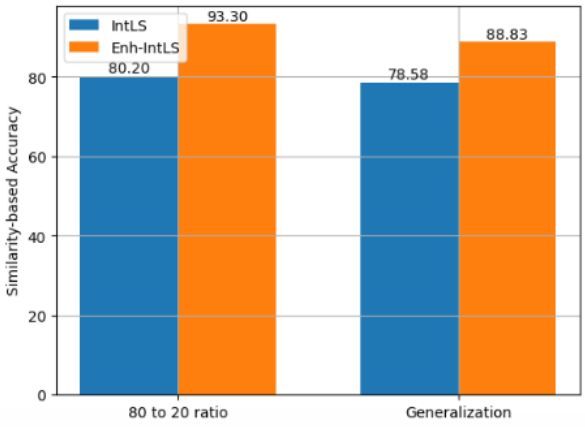


Fig. 8. Results of similarity-based accuracy metric.

As shown in Fig. 8, the difference between the baseline training-to-testing ratio and the generalization results is small (approximately 1.4% for IntLS and about 4% for Enh-IntLS). This indicates that the systems demonstrate acceptable generalization capability and can be applied in real-world environments to enhance researchers' trust in the system.

D. Accuracy, Precision, Recall, and Sensitivity Evaluation of IntLS and Enh-IntLS (Demand Prediction Task)

In this subsection, the demand prediction task is evaluated, and corresponding results are discussed. Fig. 9 and 10 illustrate confusion matrix visualizations for both systems.

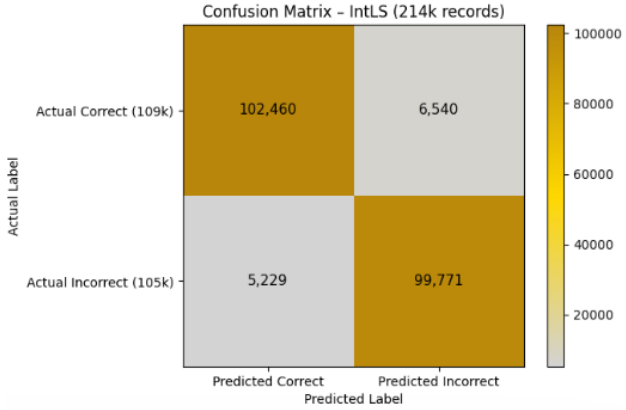


Fig. 9. Confusion matrix values of IntLS.

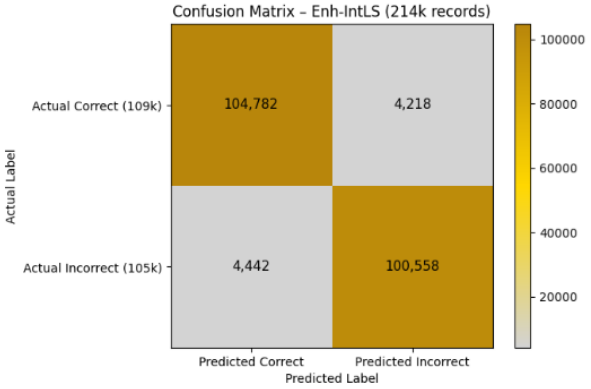


Fig. 10. Confusion matrix values of Enh-f IntLS.

Table III summarizes the obtained results without involving the generalization quality attribute.

TABLE III. RESULTS OF AI-BASED METRICS [80 TO 20] RATIO

System	Metric	AVG Value	Class	
			Correct Demand	*Incorrect Demand*
IntLS	Accuracy	94.50	92.34	96.66
	Precision	95.14	93.4	96.88
	Recall	94.00	94	94.00
	Specificity	95.02	95.02	95.02
Enh-IntLS	Accuracy	95.95	96.22	95.68
	Precision	95.93	95.89	95.97
	Recall	96.13	96.13	96.13
	Specificity	95.77	95.77	95.77

As shown in Table III, the value of each metric is presented in detail according to the correct and incorrect demand classes. It is evident that Enh-IntLS achieves better scores compared to IntLS. This improvement can be attributed to the fact that resource demand encompasses a wide spectrum of semantics that are not captured by IntLS. Since accuracy is considered one of the most important metrics in the AI domain, generalization is illustrated in Fig. 11.

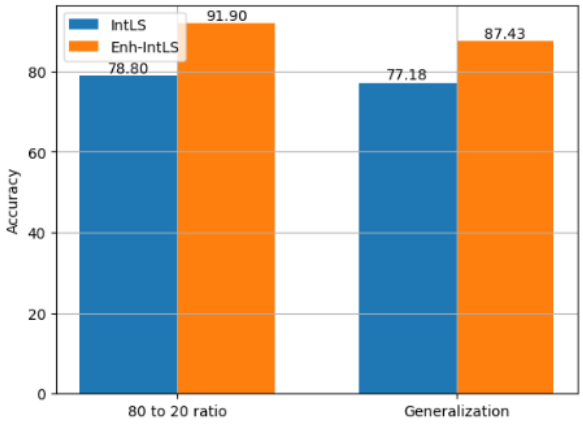


Fig. 11. Results of the accuracy metric.

As shown in Fig. 11, the difference between the baseline training-to-testing ratio and the generalization results is again small, which provides further evidence of the generalization capability of the proposed systems. However, Enh-IntLS records better values for the same reason discussed earlier when presenting the results of the similarity-based accuracy metric.

E. Results of Comparison to KNN-Based and NLP-Based Systems

In this context, only Enh-IntLS is involved in the comparison process, as it achieves better results than IntLS. For the similarity-based accuracy metric and under the generalization criterion, the previously obtained results of Enh-IntLS are considered for comparison with the selected systems. Fig. 12 illustrates the comparison results, where the KNN and NLP systems are implemented to generate a simplified version using the same selected dataset.

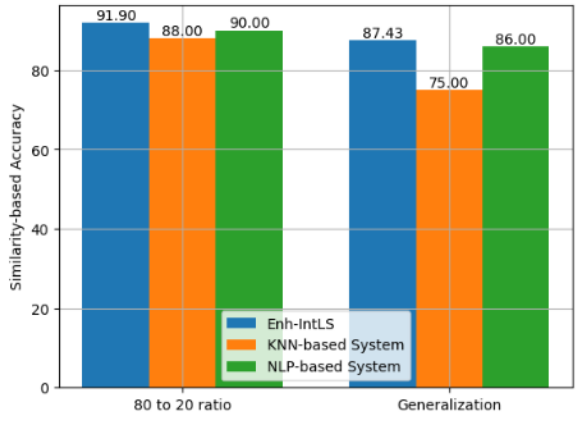


Fig. 12. Similarity-based accuracy metric comparison.

As shown in Fig. 12, the KNN-based system achieves the lowest performance compared to the other systems, as it does not incorporate knowledge graphs within its structure. This limitation leads to the retrieval of irrelevant documents within the result set. The NLP-based system achieves values close to those of the proposed system, as it employs similar NLP processing stages to those used in the development of Enh-IntLS. However, Enh-IntLS still ranks at the top because it incorporates more effective mathematical modeling within its architecture. In addition, the relationships conveyed through linking words, such as “and,” are not captured by the NLP-based system. Numerically, the Enh-IntLS achieves 1.9-based similarity-based accuracy enhancement.

Regarding the accuracy metric, which is one of the most important metrics in the AI domain, the ROC curve illustrated in Fig. 13 presents the obtained results. The ROC curve reflects the classification performance by comparing different classifiers based on the area under the curve (AUC) [16].

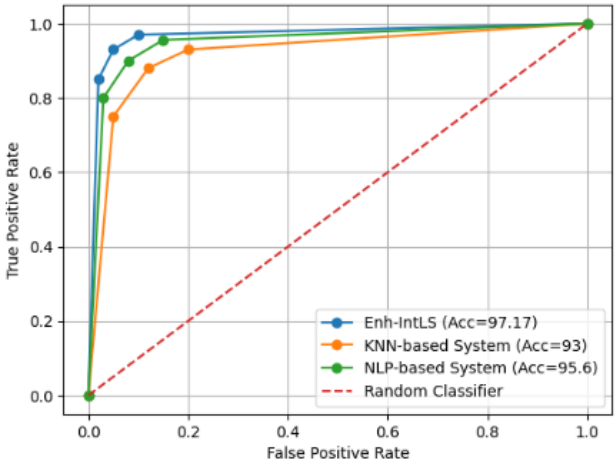


Fig. 13. ROC curve based evaluation.

As shown in Fig. 13, the Enh-IntLS system achieves the best performance, with its ROC curve located closest to the top-left corner, indicating a higher true positive rate and a lower false positive rate. The NLP-based system also demonstrates strong performance but remains slightly below the Enh-IntLS curve. In contrast, the KNN-based system shows comparatively lower performance, as its curve is farther from the optimal region. Overall, the ROC analysis confirms that Enh-IntLS provides superior discriminative capability and more reliable prediction accuracy compared to the other evaluated approaches.

V. Conclusion

This study proposed an Intelligent Library System (IntLS) designed to improve academic search and resource management by incorporating semantic understanding and predictive analytics. The system addresses key limitations of traditional library management systems, particularly the lack of semantic awareness, limited exploitation of relationships among research topics, and insufficient support for predictive decision-making. To overcome these challenges, an enhanced version of the system, Enh-IntLS, was developed by integrating knowledge graph modeling with optimized NLP

preprocessing techniques to capture hidden relationships among words and research topics. The architecture of the proposed framework combines a semantic component for generating meaningful document representations and a predictive analytics component for forecasting future resource demand based on historical usage patterns. The evaluation strategy is structured such that IntLS and Enh-IntLS are assessed using similarity-based accuracy, precision, recall, and sensitivity metrics. Subsequently, the best results obtained by Enh-IntLS are compared with two similar existing approaches to demonstrate its superiority. Experimental evaluation using the Microsoft Academic Graph dataset demonstrated that Enh-IntLS outperforms both the baseline IntLS and other comparison systems, achieving improvements in similarity-based accuracy and AI-based evaluation metrics. Furthermore, the results confirmed that the proposed systems maintain acceptable generalization capability under varying training-to-testing ratios. These findings indicate that the proposed approach can significantly enhance the effectiveness of academic library search systems and support better resource discovery and management for researchers.

Some of the non-functional requirements, as illustrative screenshots, are presented in the Appendix.

VI. Limitations and Future Work

The main limitation of this study is that computational complexity was not analyzed, and cross-validation techniques were not employed to evaluate their impact on accuracy. Furthermore, only one dataset was used to train and test the proposed system. In future work, these limitations will be addressed by providing a detailed analysis of computational complexity and by using multiple datasets for training and testing to further validate the robustness of the proposed approach. In addition, generalization is considered as internal validation, and in future work, we tend to prove generalization using real data.

Acknowledgment

The authors used ChatGPT for proofreading and generating illustrative figures. All scientific content, analysis, and conclusions were developed by the authors.

References

[1] Bagewadi, Pooja, et al. "Automating Libraries: The role of library management systems in modern library operations." International Journal of Research in Library Science 11.2 (2025): 105-114.

[2] Liu, Ze, and Bo Shao. "A systematic review of library services platforms research and research agenda." Library & Information Science Research 46.4 (2024): 101325.

[3] Khalid, Ayesha, Ghulam Farid Malik, and Khalid Mahmood. "Sustainable development challenges in libraries: A systematic literature review (2000–2020)." The Journal of academic librarianship 47.3 (2021): 102347.

[4] Kanumarlapudi, Praveen Kumar. "AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning." Learning 2.2 (2025): 1-17.

[5] Cox, Andrew M., and Sheila Corrall. "Evolving academic library specialties." Journal of the American Society for Information Science and technology 64.8 (2013): 1526-1542.

[6] Yudov, Andriy. "Traditional and modern library services: Transformation." Library Science. Record Studies. Informology 3.19 (2023): 34-41.

[7] Ex Libris Web Site. Available [Online]: https://exlibrisgroup.com/products/alma-library-services-platform/

[8] Wang, Changjun, Fengxia You, and Yu Wang. "Text intelligent classification model based on improved KNN algorithm in library service transformation." Systems and Soft Computing 7 (2025): 200313.

[9] Sawant, Pradnya, and Kavita Sonawane. "NLP-based smart decision making for business and academics." Natural Language Processing Journal 8 (2024): 100090.

[10] Aloufi, Omar F., et al. "A survey on predicting workloads and optimizing QoS in the cloud computing." 2021 International Congress of Advanced Technology and Engineering (ICOTEN). IEEE, 2021.

[11] Naveen, L., Ananda Raj, and S. Rajalakshmi. "Abstractive text summarizer: A comparative study on dot product attention and cosine similarity." 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, 2021.

[12] Xiong, Chenyan, Russell Power, and Jamie Callan. "Explicit semantic ranking for academic search via knowledge graph embedding." Proceedings of the 26th international conference on world wide web. 2017.

[13] Microsoft Academic Graph dataset Website. Available [Online]: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph (Accessed: 16/03/2026)

[14] Vimal, Bhartendoo, and S. Anupama Kumar. "Application of logistic regression in natural language processing." Int J Eng Res 9.06 (2020).

[15] Hsu, Li-Yuan, et al. "Toward building an academic search engine understanding the purposes of the matched sentences in an abstract." IEEE Access 9 (2021): 109344-109354.

[16] Janssens, A. Cecile JW, and Forike K. Martens. "Reflection on modern methods: Revisiting the area under the ROC Curve." International journal of epidemiology 49.4 (2020): 1397-1403.

Appendix

This appendix is provided to cover some of the non-functional requirements as illustrative screenshots.

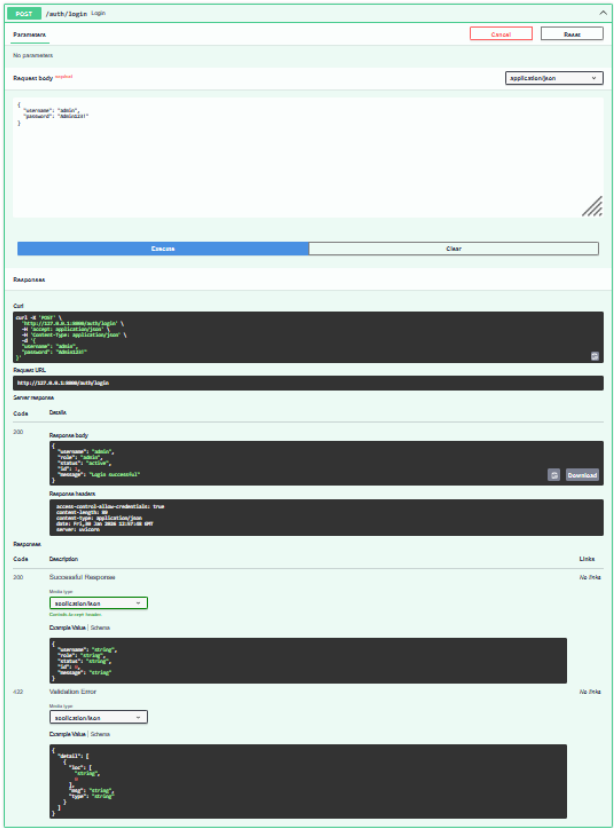


Fig. 14. Login interface.

User authentication successfully validated credentials against securely hashed password storage using bcrypt. Session management maintains user state across multiple requests without requiring repeated authentication. Role-based access control correctly restricted administrative operations to authorized users while permitting standard operations for all authenticated users. Fig. 14 demonstrates successful user authentication through the system login interface.

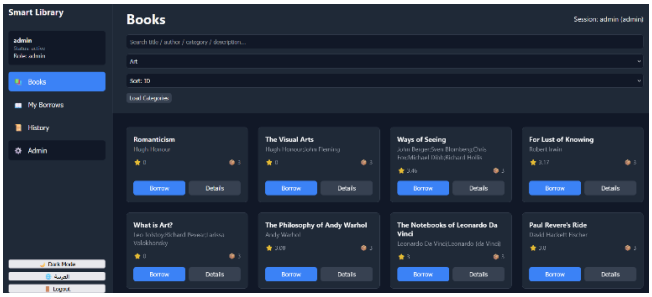


Fig. 15. Browsing interface.

The catalog search functionality correctly retrieved bibliographic records matching user-specified queries, filters, and sorting criteria. Full-text search operated across the title, author, category, and description fields. Filtering by category and sorting by relevance, title, or publication year produced the expected result sets. Fig. 15 illustrates the catalog browsing interface with active search and filter controls.

Borrowing operations correctly decremented the available copy counts and created borrowing records with appropriate due dates. The system prevented borrowing when no copies were available and enforced borrowing limits per user. Return operations restored availability, updated the borrowing status, and calculated overdue periods when applicable, as shown in Fig. 16.

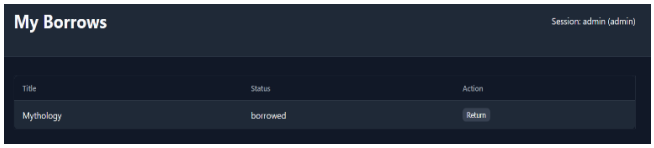


Fig. 16. Update borrowing interface.

Rice Pest Detection Using Enhanced YOLOv8n with Multilayer Contextual Attention and Deformable Snake Convolution

Shuangyuan Li[1]*, Jianglong Lin[2], Jiaming Liang[3], Tianyu Li[4]
Information Center, Jilin University of Chemical Technology, China[1]
School of Information and Control Engineering, Jilin University of Chemical Technology, China[2, 4]
Beijing Chaitin Technology Co., Ltd,Beijing, 100101, China[3]
No. 45, Chengde Street, Longtan District, Jilin 132022, P. R. China[4]

***Abstract*—Accurate and intelligent detection of rice pests is critical for ensuring food security and advancing precision agriculture. However, due to the small size, irregular morphology, dense distribution, and complex backgrounds of pest targets, traditional lightweight detection models often suffer from low recall and poor localization accuracy in real-world paddy field environments. To address these challenges, this study proposes an enhanced detection model, YOLO-Rice, based on the YOLOv8n framework. First, a Multilayer Contextual Attention (MLCA) module is embedded after the SPPF layer to collaboratively fuse channel, spatial, local, and global contextual information, thereby enhancing the model's sensitivity to subtle pest features. Second, the original C2f structure is redesigned into C2f-DS, which integrates Dynamic Snake Convolution (DSConv) to improve adaptive perception of deformable pest contours and irregular morphological edges. Finally, the conventional CIoU loss is replaced with a WIoU loss to guide the network to focus more effectively on hard-to-fit small and occluded targets. Extensive experiments on a self-constructed dataset of 11 common rice pest species demonstrate that YOLO-Rice achieves 84.8% Precision, 69.9% Recall, 78.7% mAP@0.5, and 63.4% mAP@0.5:0.95, representing significant improvements over the baseline YOLOv8n model. The proposed approach achieves an excellent balance between detection accuracy and computational efficiency, making it highly suitable for real-time deployment on UAVs and edge devices in agricultural pest monitoring applications.**

***Keywords*—Rice pest detection; YOLOv8n; deep learning; real-time detection**

I. Introduction

In recent years, with the continuous advancement of deep learning technology, image-based intelligent detection methods have become an essential pathway to ensure food security and promote agricultural modernization [1]. As one of the world's most important staple crops, rice is highly susceptible to a variety of pest infestations during its growth cycle, including Delphacidae, Cicadellidae, Crambidae, and Curculionidae. These pests not only hinder the healthy growth of rice but can also lead to severe yield reduction or even total crop failure, posing a significant threat to national food security. Consequently, early identification and precise localization of rice pests have become critical research focuses in modern agricultural pest management [2].

Traditional pest monitoring approaches, such as manual inspection and trap-based surveillance, suffer from high labor intensity, low efficiency, and subjective bias, making them unsuitable for large-scale field monitoring. With the rapid development of unmanned aerial vehicle (UAV) remote sensing and the increasing availability of agricultural imagery, computer vision – based automatic pest detection has emerged as a promising solution. Among numerous object detection frameworks, the YOLO (You Only Look Once) series has gained significant attention due to its end-to-end architecture, fast inference speed, and high detection accuracy, making it particularly suitable for agricultural visual tasks[3-4].

As the latest member of the YOLO family, YOLOv8 introduces an anchor-free mechanism and a lightweight C2f structure, achieving a better balance between precision and computational efficiency. However, when applied directly to rice pest detection[5], it still faces several challenges: 1) The pest targets are generally small, densely distributed, and often confused with complex backgrounds such as veins or shadowed regions, leading to false detections and missed targets; 2) Pest bodies exhibit irregular shapes, occlusion, and deformation, making it difficult for standard convolutions to accurately capture boundary features[6]; 3) The conventional bounding box regression loss functions, such as CIoU, show insufficient adaptability to hard samples and slow convergence during training, which limits the final detection accuracy. To overcome these challenges, this study proposes a structure- and loss-enhanced detection model based on YOLOv8n, specifically designed for high-precision detection of multiple rice pest species under complex field environments. The main contributions of this work are summarized as follows:

A Multilayer Contextual Attention (MLCA) mechanism is introduced after the SPPF module to integrate spatial, channel, local, and global contextual features. This enhances the model's focus on pest regions and improves its ability to recognize small and occluded targets. Compared to conventional attention modules such as SE and CBAM, MLCA achieves superior feature discrimination without adding significant computational overhead [7].

A redesigned C2f-DS module based on Dynamic Snake Convolution (DSConv) is proposed. By introducing dynamic offset learning, DSConv adaptively adjusts convolution

*Corresponding author.

sampling positions and shapes, enabling "snake-like" contour perception. This improves the model's capability to capture irregular pest boundaries and deformable structures, particularly for non-rigid shapes such as the feeding edges of Cnaphalocrocis medinalis.

The Weighted IoU (WIoU) loss function is employed to replace the traditional CIoU, enhancing the model's robustness in bounding-box regression. WIoU dynamically assigns gradient weights to samples with varying difficulty, encouraging the network to focus more on hard-to-fit or occluded targets, thereby improving convergence stability and overall mAP performance.

To validate the proposed approach, a large-scale rice pest dataset containing 11 common pest species was constructed. Experimental results demonstrate that the improved YOLOv8n model achieves significantly higher detection accuracy and robustness in complex, cluttered, and deformable pest scenarios while maintaining lightweight deployment capability. Notably, the proposed model outperforms mainstream detectors such as YOLOv5, YOLOv7, and YOLOv11 in key metrics including mAP, Precision, and Recall, verifying its effectiveness and practical potential for real-world rice pest detection tasks.

In summary, this study provides a novel structural optimization approach for object detection in agricultural vision applications and offers a feasible technical foundation for the intelligent deployment of rice pest monitoring systems. Future research will focus on improving the model's cross-regional generalization, multimodal data fusion, and extreme lightweight deployment on agricultural embedded platforms, thereby advancing pest control toward greater efficiency, intelligence, and precision.

II. Related Research

In recent years, with the deep integration of unmanned aerial vehicles (UAVs) and deep learning technologies in precision agriculture, vision-based rice pest detection has become a key research focus. Object detection techniques have evolved from early two-stage frameworks to more efficient single-stage and anchor-free approaches, with researchers exploring model lightweighting, small-object enhancement, deformable feature extraction, and regression loss optimization. Early two-stage detectors such as Faster R-CNN [8] and Mask R-CNN [9] demonstrated strong performance in locating crop diseases and large pest regions. For instance, Ren et al. [10] combined an RPN with a fine classifier for wheat leaf spot detection, achieving 67% mAP, though its inference delay of over 200 ms limited UAV-based real-time deployment. He et al. [11] improved feature alignment through ROI-Align, significantly enhancing lesion boundary segmentation accuracy, but the method remained computationally expensive due to dense proposal generation.

To balance speed and precision, single-stage detectors have gained prominence. Liu et al. [12] introduced SSD, which enables end-to-end detection using multi-scale feature layers and achieves 59 FPS inference. However, it exhibits high miss rates for small pests such as Cnaphalocrocis medinalis and Delphacidae, whose body sizes span only tens of pixels. Lin et al. [13] proposed RetinaNet with Focal Loss to reduce gradient dominance by easy samples, significantly improving small-object recall. Redmon and Farhadi [14] developed YOLOv3, which leverages multi-scale pyramid prediction and efficient convolution to reach 45 FPS, showing robust performance in heavily occluded rice leaf environments. Later, Bochkovskiy et al. [15] released YOLOv4, integrating CSPDarkNet and PANet to better balance speed and accuracy. Jocher et al. [16] further optimized YOLOv5 with automatic anchor learning, depthwise separable convolution, and channel pruning, reducing parameters below 8M. Studies applying YOLOv5 in agricultural pest detection incorporated ECA attention and BiFPN fusion, achieving a 4 – 6% increase in mAP. Li et al. [17] developed a lightweight version of YOLOv7 for Noctuidae detection, reducing model size by 30% and enabling 30 FPS real-time inference on UAV platforms; however, detection of dense Crambidae larvae in rice fields remained unsatisfactory. More recently, Fan et al. [18] embedded SimAM attention and NWD loss into the YOLOv8 backbone and neck for grape leaf disease detection, achieving a 90.4% mAP@0.5 and demonstrating superior fine-grained focus. Nevertheless, their approach did not specifically address pest morphology or leaf occlusion, nor was it validated on diverse rice pest datasets.

Addressing the challenges of small, irregular, and occluded pest targets in paddy fields, Dai et al. [19] proposed Deformable ConvNets, which allow convolutional sampling points to adaptively shift according to object contours, thereby improving feature alignment under occlusion and deformation. Zhu et al. [20] further enhanced this concept through the Modulated DCNv2, incorporating modulation mechanisms to refine sampling precision. In addition, bounding box regression losses have been shown to significantly affect localization accuracy for small objects. While the traditional IoU loss focuses only on overlap, it fails to account for center distance and aspect ratio. To address this, Rezatofighi et al. [21] proposed GIoU, and Zheng et al. [22] extended it to DIoU and CIoU, incorporating geometric penalties that improved convergence speed and localization precision. Building upon this, Bao et al. [23] introduced WIoU v3, which dynamically weights samples according to quality, emphasizing harder instances during training. This approach improved mAP@0.5:0.95 by 4.1% in dense rice pest detection scenarios.

Despite significant progress in lightweight architectures, attention mechanisms, deformable convolutions, and regression loss optimization, most existing studies focus on isolated module improvements. A unified framework that simultaneously addresses small target sensitivity, irregular shape modeling, and robust bounding box regression—while maintaining real-time performance—is still lacking. To bridge this gap, this study integrates Multilayer Contextual Attention (MLCA), Dynamic Snake Convolution (DSConv), and WIoU loss into the anchor-free YOLOv8n architecture. This comprehensive approach aims to achieve superior detection accuracy and robustness for rice pest detection under complex field conditions, enabling efficient deployment on UAV and edge platforms.

III. Proposed Method

This study focuses on the task of rice pest detection based on YOLOv8n, addressing challenges such as small target size,

weak texture features, and irregular pest contours. Based on the original framework, the following improvements are proposed: Firstly, Multilayer Contextual Attention (MLCA) is embedded after the SPPF module to fuse channel, spatial, local, and global contextual features, enhancing the model's responsiveness to weak textures in pest regions [24]. Secondly, the C2f-DS module is designed by replacing some standard convolutions in the original C2f structure with Dynamic Snake Convolution (DSConv), which adaptively learns offsets to capture irregular edges and deformable pest contours. Finally, Weighted IoU (WIoU) loss is adopted to replace CIoU, dynamically assigning gradient weights to samples of varying difficulty, thereby improving the robustness and convergence speed of bounding box regression for small pest targets.

While maintaining the lightweight advantages of YOLOv8n, these improvements jointly optimize the model's perception and localization of "small, dense, and weak" pest targets without significantly increasing computational cost. The overall framework of the proposed YOLO-Rice model is shown in Fig. 1, and the detailed design of each module is described in the following subsections.

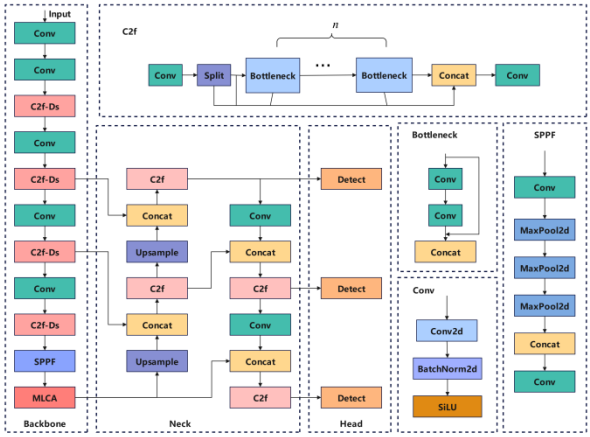


Fig. 1. Overall architecture of the proposed YOLO-Rice model.

A. Reconstruction of the Backbone Network

Rice pest targets are often extremely small and visually similar to the leaf background in both color and texture. Traditional convolutional operations struggle to distinguish pests from background interference, leading to frequent false detections and missed detections. Therefore, it is essential to enhance the model's focus on pest regions through an attention mechanism while maintaining lightweight efficiency. Existing attention modules, such as SE and CBAM, perform one-dimensional weighting along either the channel or spatial axis, which can result in the loss of local information for small targets [25].

To address this limitation, a Multilayer Contextual Attention (MLCA) mechanism is proposed and embedded after the SPPF module. The improved module fuses local pooling, global pooling, and one-dimensional convolution to achieve bidirectional attention across channel and spatial dimensions while integrating both local and global contexts. This structure provides more fine-grained feature enhancement, strengthening the network's feature extraction capability. The structure of MLCA is shown in Fig. 2.

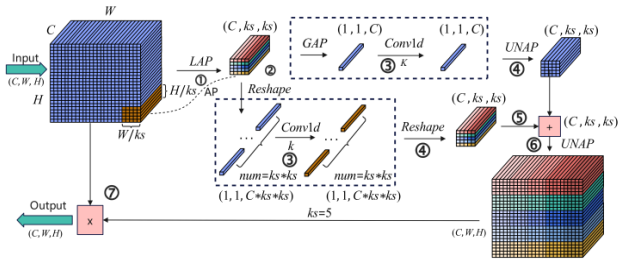


Fig. 2. MLCA structure.

As illustrated in Fig. 2, the MLCA module achieves fine-grained pest feature reinforcement through two collaborative dimensions: local – global complementarity and channel – spatial interaction. Specifically, the input feature map is first processed by Local Average Pooling (LAP) to extract coarse local features from subdivided sub-blocks. It is then divided into two branches: a global branch that applies Global Average Pooling (GAP) to highlight overall field-level context, and a local branch that reshapes the LAP output to preserve channel-level details within each subregion. Both branches are passed through 1D convolution (Conv1d) layers for channel compression and feature transformation. Subsequently, the features are restored to the original spatial resolution through reshaping and upsampling via Unpooled Average Pooling (UNAP), and the two branches are fused along the channel dimension to achieve semantic – detail interaction. Finally, the fused result is expanded to the input resolution and multiplied element-wise with the original feature map to generate a weighted feature map with stronger pest region responses. The functions of the key operations in the MLCA module are explained in Table I.

TABLE I. EXPLANATION OF KEY OPERATIONS IN THE MLCA MODULE

Operation	Role	Output
LAP	Preserve fine-grained spatial details (e.g., pest edges, feeding marks)	Local texture features
GAP	Compresses global context (e.g., leaf-background separation)	Global channel descriptor
UNAP	Aggregates neighborhood information to bridge local and global scales	Medium-scale context
Conv1D	Models inter-channel dependencies (kernel size=5)	Channel attention weights
Feature Fusion	Sums LAP, GAP, UNAP, then multiplies by the original feature map	Refined multi-scale features

To demonstrate the lightweight efficiency of MLCA, it is compared with SE and CBAM in terms of additional parameters, FLOPs, and mAP improvement. The results in Table II show that MLCA introduces only 0.12M parameters and 0.48 GFLOPs, significantly lower than CBAM, while achieving the highest mAP gain (+3.7% mAP@0.5). This confirms that MLCA provides superior feature discrimination with minimal computational overhead, making it ideal for real-time UAV deployment.

TABLE II. COMPARISON OF ATTENTION MODULES ON YOLOV8N BASELINE

Attention Module	Parameters (M)	FLOPs (G)	mAP@0.5 (%)	Δ mAP@0.5 (%)
None (baseline)	5.30	8.1	70.4	-
SE	5.41 (+0.11)	8.2	72.9	+2.5
CBAM	5.45 (+0.15)	8.3	74.2	+3.8
MLCA	5.42 (+0.12)	8.3	74.1	+3.7

B. C2f-DS Module with Dynamic Snake Convolution

Rice pests often exhibit non-rigid, elongated, or curved morphological structures. For instance, Crambidae larvae create irregular chewing marks along leaf edges; Cicadellidae closely resemble leaf textures and are easily overlooked; and Curculionidae often appear partially occluded due to their shrimp-like shapes. Traditional convolutions with fixed sampling points cannot adapt to such boundary variations, leading to decreased detection accuracy.

To overcome these limitations, Dynamic Snake Convolution (DSConv)—a deformable convolution variant capable of adaptively focusing on elongated and curved structures—is integrated into the C2f framework, resulting in a novel module termed C2f-DS. The architecture of DSConv is shown in Fig. 3.

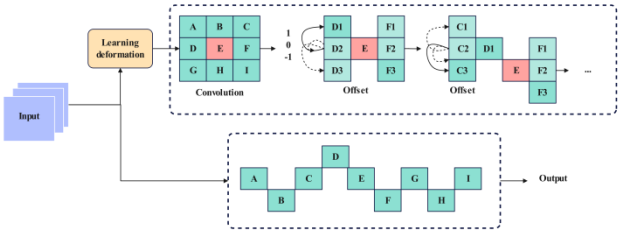


Fig. 3. Architecture of dynamic snake convolution.

The C2f module in YOLOv8n is a lightweight feature extraction unit that balances computational efficiency and representational power. DSConv is embedded into C2f rather than the backbone's basic convolutions or the neck for the following reasons. Firstly, feature hierarchy: C2f operates at multiple scales within the backbone, capturing both fine-grained details and semantic information. Embedding DSConv in C2f allows deformable sampling to be applied across a range of receptive fields, which is essential for detecting pests of varying sizes and deformations. Secondly, Computational efficiency: Replacing all backbone convolutions with DSConv would significantly increase parameters and inference time. C2f already contains a split-and-merge design; replacing only a portion of its standard convolutions with DSConv minimizes additional cost while preserving the benefits of deformable convolution. Finally, Residual integration: The C2f structure inherently includes residual connections, which help stabilize training when introducing dynamic offsets. This mitigates the risk of gradient oscillation caused by excessive deformation [26].

As illustrated in Fig. 4, the proposed C2f-DS module introduces an iterative "snake-like" offset mechanism based on deformable convolution and deeply integrates it within the original C2f structure. This design improves feature extraction while maintaining computational efficiency. Specifically, C2f-DS dynamically samples local pest regions, adaptively capturing irregular leaf-edge contours caused by Crambidae larvae and subtle differences between Cicadellidae nymphs and leaf veins. This significantly enhances the model's ability to fit small pest boundaries. Unlike stacking multiple deformable convolutions, C2f-DS replaces only part of the standard convolutional kernels in C2f, preserving the residual and channel-split design. As a result, the number of parameters and computational cost remain close to those of YOLOv8n, with negligible impact on inference speed.

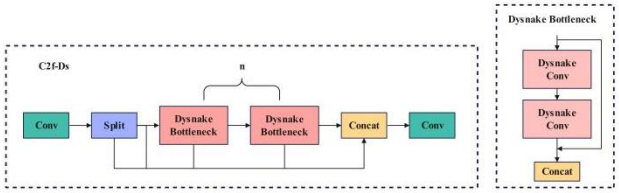


Fig. 4. Structure of the C2f-DS module.

TABLE III. COMPARISON BETWEEN DSCONV (C2F-DS) AND STANDARD CONVOLUTION

Aspect	Standard Convolution	DSConv (C2f-DS)
Sampling positions	Fixed grid, axis-aligned	Adaptive offsets, distributed along target contours
Adaptability to pest boundaries	Low – cannot fit curved or deformed edges	High – fits irregular contours of Crambidae, Curculionidae, etc.
Offset learning	Not applicable	Learned via lightweight auxiliary convolution (~3×3 depthwise)
Parameter increase	Baseline	<5% increase compared to standard C2f
Inference speed impact	Baseline	Comparable (~1–3 FPS drop on RTX 3060)
Best suited for	General feature extraction	Small, elongated, occluded, or clearly deformed pest targets

Unlike standard convolution with a fixed sampling grid, DSConv enhances the perception of irregular pest boundaries by learning adaptive offsets. Specifically, a lightweight auxiliary convolutional layer (typically a 3×3 depthwise convolution) takes the input feature map and predicts a 2D offset field $\triangle p_n$ for each sampling point p_n of the kernel. During training, these offsets are jointly optimized with the main network weights via backpropagation; bilinear interpolation is used to ensure differentiability of fractional sampling positions. At inference, the kernel samples features at shifted positions $p_n + \Delta p_n$, allowing the receptive field to "snake" along elongated or curved pest contours (e.g., the feeding edge of Cnaphalocrocis medialis larvae). Table III compares the key differences between DSConv in the C2f-DS module and standard convolution. It can be seen that DSConv significantly improves boundary fitting for deformable targets while maintaining a lightweight design.

Furthermore, a residual connection is introduced in the iterative offset process to improve training stability and mitigate gradient oscillations caused by excessive deformation. This module significantly enhances detection robustness and precision for small rice pests while maintaining a lightweight structure, making it well-suited for real-time UAV-based agricultural applications.

C. Weighted IoU (WIoU) Loss Function

In small-object pest detection, the precision of bounding box regression plays a crucial role in determining overall detection performance. The original YOLOv8n employs the Complete IoU (CIoU) loss, which simultaneously optimizes IoU, center distance, and aspect ratio. However, for extremely small, irregularly shaped, and background-blended pest targets, CIoU tends to overemphasize high-IoU samples, making it less effective for low-IoU or fuzzy-boundary samples. This often results in inaccurate bounding box regression and higher miss rates [27].

To address this issue, CIoU is replaced with the Weighted IoU (WIoU) loss function. WIoU introduces a dynamic focusing mechanism that adjusts gradient weights according to prediction quality, enabling the model to pay more attention to hard-to-fit small pest boundaries. The relationship between the ground truth and predicted bounding boxes is illustrated in Fig. 5.

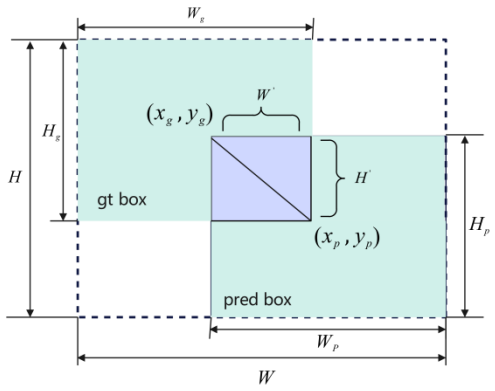


Fig. 5. Illustration of ground truth box and predicted box.

The key mathematical difference between CIoU and WIoU lies in how gradient weights are assigned. CIoU uses a fixed penalty term based on center distance and aspect ratio, treating all samples uniformly. In contrast, WIoU introduces a dynamic focusing coefficient that reduces the contribution of easy samples (high IoU) and amplifies the gradient for hard samples (low IoU). The WIoU loss is calculated as shown in Eq. (1)-(3).

$$L_{WIOU} = r \cdot L_{IOU} \tag{1}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{2}$$

$$\beta = \frac{(W - W_g)^2 + (H - H_g)^2}{{W_g}^2 + {H_g}^2} \tag{3}$$

where, β is the outlier degree, and α, δ are hyperparameters controlling the focusing strength. The weighting factor r dynamically adjusts the gradient magnitude: for samples with small β (easy samples, high IoU), r is small, reducing their contribution; for samples with large β (hard samples, low IoU), r approaches 1, emphasizing the loss. This mechanism forces the network to prioritize poorly localized small targets, resulting in more precise bounding box predictions, especially for occluded or irregularly shaped pests.

IV. Experiments and Analysis

A. Dataset and Experimental Setup

This study targets 11 common rice pest species: Curculionidae, Delphacidae, Cicadellidae, Phlaeothripidae, Cecidomyiidae, Hesperiidae, Crambidae, Chloropidae, Ephydridae, Noctuidae, and Thripidae. The dataset contains 4,559 pest instances manually annotated in YOLO format using LabelImg. The images were split into training, validation, and test sets at a ratio of 7:2:1, resulting in 3,191, 911, and 457 images, respectively, while maintaining class balance across subsets. As shown in Fig. 6, Delphacidae and Cicadellidae dominate the dataset, while Thripidae and Noctuidae are underrepresented, exhibiting a clear long-tail distribution. The bounding box widths and heights are mostly concentrated in small-scale regions (normalized values between 0.0 and 0.4), indicating that most pest targets are small and morphologically diverse.

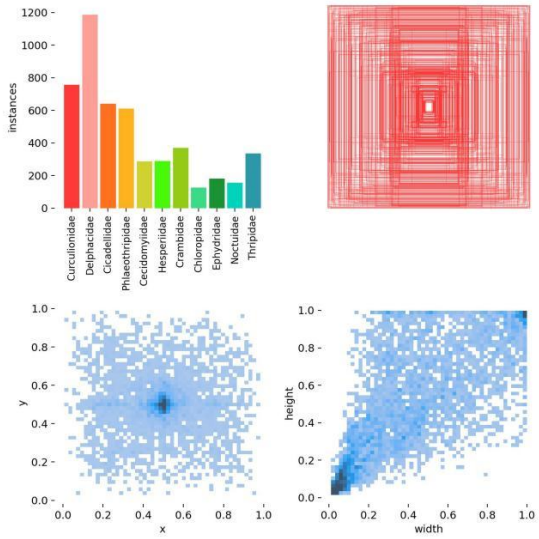


Fig. 6. Distribution of target categories and sizes in the training set.

Partial samples of rice pests are shown in Fig. 7.



Fig. 7. Partial samples of rice pests.

All experiments were conducted on a Windows 11 system with an Intel Core i9-13900H CPU and an NVIDIA GeForce RTX 3060 GPU. Training was accelerated using PyTorch 2.1.0 with CUDA 11.8. All models were trained under identical hyperparameters: the SGD optimizer with 200 epochs, a batch size of 64, and an initial learning rate of 0.01.

B. Evaluation Metrics

To comprehensively evaluate the proposed method, three key metrics are employed: Precision (P), Recall (R), and Mean Average Precision (mAP) to validate the detection accuracy, while Parameter Count and Frame Rate (FPS) assess the model compactness and speed. These metrics are defined as shown in Eq. (4)-(7).

$$P = \frac{TP}{(TP + FP)} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{c} AP_i}{c} \tag{6}$$

$$FPS = \frac{1000ms}{(t1 + t2 + t3)} \tag{7}$$

Here, TP denotes the number of correctly predicted positive samples, FP indicates the number of incorrectly predicted negative samples, FN represents the number of missing targets among the correct ones, c stands for the number of defect categories, t1 is the image preprocessing time, t2 is the image inference time, and t3 is the post-processing time.

C. Ablation Experiments

To verify the effectiveness of the proposed Multilayer Contextual Attention (MLCA), C2f-DS module, and Weighted IoU (WIoU) loss function in the rice pest detection task, a series of ablation experiments were designed and conducted. All experiments were based on the YOLOv8n architecture as the baseline model. Under identical dataset partitioning, data augmentation strategies, and hyperparameter configurations, each improvement module was progressively introduced to quantitatively evaluate the individual contributions and synergistic effects. The experimental results are shown in Table IV.

TABLE IV. ABLATION EXPERIMENT RESULTS

Configuration	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters (M)
Baseline	0.800	0.632	0.704	0.562	5.3
+ MLCA	0.823	0.658	0.741	0.595	5.4
+ C2f-DS	0.816	0.652	0.734	0.588	6.8
+ WIoU	0.812	0.645	0.729	0.580	5.3
MLCA + C2f-DS	0.836	0.678	0.762	0.612	6.9
MLCA + WIoU	0.828	0.665	0.752	0.603	5.4
C2f-DS + WIoU	0.825	0.662	0.748	0.599	6.8
YOLO-Rice	0.848	0.699	0.787	0.634	7.8

From the experimental results in Table IV, it can be observed that each improvement module effectively enhances model performance. When introduced individually, MLCA, C2f-DS, and WIoU improve mAP@0.5 by 3.7%, 3.0%, and 2.5%, respectively, validating their independent effectiveness in feature enhancement, deformable modeling, and regression optimization. In terms of module combinations, the combination of MLCA and C2f-DS achieves the most prominent synergistic effect, with mAP@0.5 reaching 0.762, representing improvements of 2.1% and 2.8% compared to using each module alone, indicating a significant complementary effect between the attention mechanism and deformable convolution. The combinations of MLCA with WIoU and C2f-DS with WIoU also demonstrate favorable synergistic effects. Finally, the YOLO-Rice model, which fully integrates all three modules, achieves optimal performance across all evaluation metrics, with Precision, Recall, mAP@0.5, and mAP@0.5:0.95 reaching 0.848, 0.699, 0.787, and 0.634, respectively, representing improvements of 4.8%, 6.7%, 8.3%, and 7.2% compared to the baseline YOLOv8n. Although the number of parameters increases from 5.3M to 7.8M, the model still maintains a real-time inference speed of 62 FPS, meeting the deployment requirements for UAVs and edge devices. In summary, the ablation experiments fully validate the individual effectiveness and combined synergy of the MLCA, C2f-DS, and WIoU modules in the rice pest detection task. The three modules work together from three dimensions—feature enhancement, deformable modeling, and regression optimization—forming a comprehensive detection optimization system.

D. Comparison of Dysnake Conv Embedding Across Different C2f Layers

To further investigate the influence of DSConv placement within the backbone, embedding DSConv into C2f_1 (shallow), C2f_2 (middle), C2f_3 (deep), and C2f_4, as well as combinations thereof, was compared against the baseline model with standard convolutions. Only the C2f structure at the specified layers was modified; all other configurations remained unchanged. The results are shown in Table V.

TABLE V. EXPERIMENTAL RESULTS OF DSCONV EMBEDDING AT DIFFERENT C2F LAYERS

Dysnake Conv Embed location	Precision	Recall	mAP@0.5	mAP@0.5：0.95	Parameters (M)
No replacement	0.800	0.632	0.704	0.562	5.3
C2f_1	0.810	0.648	0.718	0.576	5.6
C2f_2	0.816	0.652	0.734	0.588	5.8
C2f_3	0.808	0.644	0.725	0.581	6.8
C2f_4	0.803	0.640	0.717	0.573	6.1
C2f_2+C2f_3	0.823	0.662	0.745	0.599	6.5
C2f_1 + C2f_2 + C2f_3	0.829	0.668	0.752	0.605	7.0
C2f_3 + C2f_4	0.819	0.656	0.740	0.593	6.3
Replace All	0.833	0.670	0.760	0.610	7.4

The results indicate that applying DSConv across all four C2f layers significantly enhances multi-scale feature extraction—from fine-grained leaf textures to overall pest semantics—especially for small, deformed, and partially occluded pests such as Crambidae and Delphacidae. Compared to replacing only middle and deep layers, the full-layer configuration increased mAP@0.5 to 0.760 (+1.5%) and mAP@0.5:0.95 to 0.610 (+1.1%), while Precision and Recall reached 0.833 and 0.670, respectively. Although this setup increased parameters by 2.1M, the detection precision and robustness in complex rice-field backgrounds improved significantly, validating the advantage of multi-level collaborative modeling of deformable pest features.

E. Comparison of Attention Mechanisms

This section compares the proposed MLCA with other mainstream attention mechanisms, including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA). All models were built upon YOLOv8n, differing only in the attention module applied. Training configurations were identical across all experiments. The results are presented in Table VI.

TABLE VI. COMPARISON OF ATTENTION MECHANISMS

Model Version	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Baseline_NoAttn	0.800	0.632	0.704	0.562
SE_Attn	0.812	0.645	0.729	0.579
CA_Attn	0.818	0.650	0.735	0.586
CBAM_Attn	0.821	0.654	0.742	0.592
MLCA_Attn	0.823	0.658	0.741	0.595

As shown in Table VI, SE improved mAP@0.5 from 0.704 to 0.729 and mAP@0.5:0.95 from 0.563 to 0.579, demonstrating its effectiveness in suppressing background noise. CA, which captures long-range spatial dependencies, further improved mAP@0.5 to 0.735 and mAP@0.5:0.95 to 0.586. CBAM, integrating both channel and spatial attention, yielded mAP@0.5=0.742 and mAP@0.5:0.95=0.592. The proposed MLCA achieved the highest overall performance, with mAP@0.5:0.95=0.595, Precision=0.823, and Recall=0.658. These results highlight that MLCA provides stronger robustness and localization accuracy by jointly modeling weak pest details and global contextual semantics.

F. Comparison of Loss Functions

This section evaluates the impact of different bounding box regression loss functions. The default CIoU loss in YOLOv8n was replaced with GIoU, DIoU, and WIoU, while maintaining identical training settings. The results are shown in Table VII.

TABLE VII. COMPARISON OF LOSS FUNCTIONS

Model Version	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Baseline_CIoU	0.800	0.632	0.704	0.562
GIoU_Loss	0.801	0.628	0.701	0.553
DIoU_Loss	0.806	0.635	0.710	0.559
WIoU_Loss	0.812	0.645	0.729	0.580

The results show that incorporating center distance and aspect ratio constraints (DIoU, CIoU) improves regression precision and convergence compared to IoU and GIoU. Building upon this, WIoU introduces sample difficulty-aware weighting, allowing the model to focus more on hard-to-detect small or occluded pests. Compared with CIoU, WIoU achieved improvements across all metrics: Precision increased from 0.800 to 0.812, Recall from 0.632 to 0.645, mAP@0.5 by 2.5%, and

mAP@0.5:0.95 by 1.8%. These results confirm that WIoU enhances the alignment between predicted and ground-truth boxes while effectively suppressing training noise from anomalous samples, yielding more stable and generalized detection across multiple IoU thresholds.

G. Comparison with State-of-the-Art Models

To comprehensively evaluate the overall performance of the proposed YOLO-Rice model in the rice pest detection task, this section compares it with several mainstream object detection algorithms. All compared models were trained and tested under identical dataset partitioning, data augmentation strategies, and hyperparameter configurations to ensure fairness. The experimental results are shown in Table VIII.

As shown in Table VIII, YOLO-Rice achieves the best detection performance among all compared models. As traditional detectors, Faster R-CNN and SSD achieve mAP@0.5 of 0.668 and 0.652, respectively, but their large parameter sizes and low inference speeds make them unsuitable for real-time UAV monitoring. Among the lightweight YOLO series models, YOLOv5s, YOLOv7, YOLOv8n, YOLOv9, YOLOv10n, and YOLOv11n all demonstrate a good balance between accuracy and speed. Notably, YOLOv11n achieves a mAP@0.5 of 0.740 and an inference speed of 150 FPS with only 5.5M parameters, exhibiting excellent lightweight performance. Building upon this, the proposed YOLO-Rice model achieves further optimization, with mAP@0.5 reaching 0.787, representing improvements of 8.3% over the baseline YOLOv8n and 4.7% over the latest YOLOv11n. The mAP@0.5:0.95 reaches 0.634, an improvement of 7.2% over YOLOv8n. Precision and Recall reach 0.848 and 0.699, respectively, representing increases of 4.8% and 6.7% compared to YOLOv8n. In terms of inference speed, YOLO-Rice achieves 138 FPS on an RTX 4090, far exceeding the minimum frame rate requirement for real-time detection, fully validating its ability to achieve significant detection accuracy improvements while maintaining lightweight characteristics.

TABLE VIII. PERFORMANCE COMPARISON OF DIFFERENT DETECTION MODELS ON THE RICE PEST DATASET

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS (4090)	Params (M)
Faster R-CNN	0.756	0.582	0.668	0.512	42	41.5
SSD (VGG16)	0.742	0.568	0.652	0.498	85	26.3
YOLOv5s	0.780	0.600	0.680	0.535	156	7.2
YOLOv7	0.790	0.620	0.700	0.552	128	36.9
YOLOv8n	0.800	0.632	0.704	0.562	142	5.3
YOLOv9	0.818	0.655	0.726	0.578	135	8.9
YOLOv10n	0.815	0.655	0.725	0.576	148	5.7
YOLOv11n	0.825	0.670	0.740	0.589	150	5.5
YOLO-Rice	0.848	0.699	0.787	0.634	138	7.8

H. Confusion Matrix Analysis

To further analyze the classification performance of the YOLO-Rice model across different rice pest categories, this section presents the normalized confusion matrices, as shown in Fig. 8. Rows of the confusion matrix represent ground-truth labels, columns denote model predictions, and diagonal elements indicate the correct classification ratios for each category, while off-diagonal elements reflect misclassification cases. The confusion matrix enables intuitive evaluation of the model's recognition capability for individual pest categories and the degree of confusion between classes.

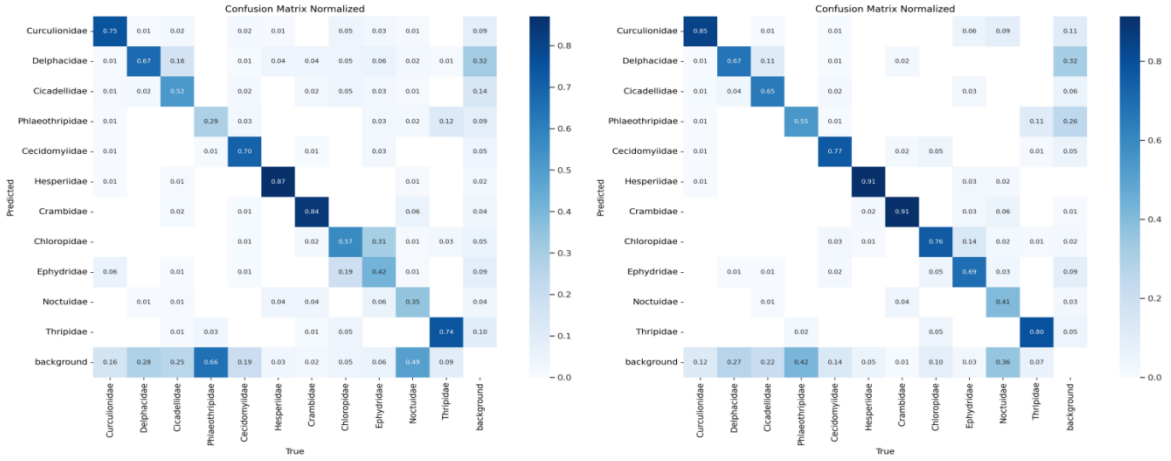


Fig. 8. Confusion matrix: YOLOv8 (left) and YOLO-Rice (right).

As observed from Fig. 8, YOLO-Rice achieves a significant improvement in overall classification accuracy, with substantial increases in the recall rates of most categories. Specifically, the recall for Curculionidae rises from 0.75 to 0.85, for Phlaeothripidae from 0.29 to 0.55, for Ephydridae from 0.42 to 0.69, for Hesperiidae and Crambidae both reaching 0.91, and for Thripidae up to 0.80. Only Delphacidae remains unchanged at 0.67, and Noctuidae sees a slight increase to 0.41. These results demonstrate that the model significantly enhances its ability to identify difficult-to-classify samples and categories with similar morphological features, while also reducing inter-class confusion notably—for instance, the confusion between Ephydridae and Chloropiidae decreases from 0.31 to 0.05. The sole exception is the background class, whose recall drops from 0.66 to 0.42. This decline primarily arises from the model's increased focus on recognizing insect categories, leading to exacerbated misclassification of background regions (e.g., leaf shadows, water reflections, dead spots) as pest taxa such as Phlaeothripidae. While this trade-off is acceptable in some practical applications, it still calls for further optimization to reduce false positives. To address this issue, future work will adopt three strategies: 1) introducing a background suppression loss term that penalizes false positives of the background class; 2) augmenting the training set with diverse negative samples (e.g., empty paddy field images under various lighting conditions); and 3) applying class-balanced sampling or a variant of Focal Loss to down-weight easily confused background patterns. These measures aim to maintain high recall for pest targets while effectively controlling background misclassification, thereby improving overall robustness in real-world paddy field environments. Overall, the proposed YOLO-Rice model, through optimized feature extraction, network architecture, and data augmentation, effectively improves the accuracy and robustness of insect family classification, and the above-mentioned strategies will further enhance its practical deployability.

I. Visualization Analysis of Detection Results

To intuitively demonstrate the practical performance of the proposed YOLO-Rice algorithm in rice pest detection tasks, this study selects detection results from various complex field scenarios for visual comparison and analysis. The Precision–Recall curves before and after improvement are shown in Fig. 9, and the visualization results are presented in Fig. 10.

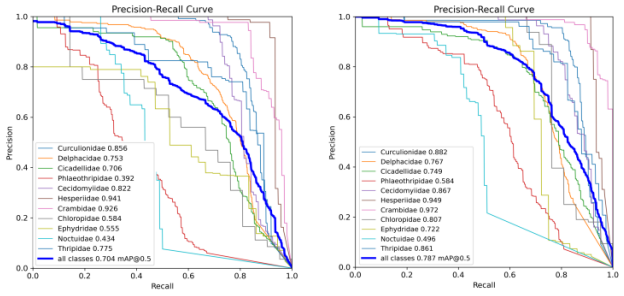


Fig. 9. Precision-recall curve.

Fig. 9 presents the Precision-Recall (PR) curves of the baseline YOLOv8n and the proposed YOLO-Rice model across 11 categories of rice pests. The proposed model achieves a significant overall improvement, with mAP@0.5 increasing from 0.704 to 0.787 (+8.3%). Notably, recognition accuracy for Noctuidae, Delphacidae, and Cicadellidae—categories typically characterized by small size and high occlusion—exhibits remarkable gains of 5.2%, 1.4%, and 4.3%, respectively. This improvement demonstrates that the integration of MLCA and DSConv effectively enhances the model's capacity to distinguish fine-grained pest textures and irregular contours under complex paddy field conditions. The optimized WIoU loss further contributes to improved bounding box regression accuracy, particularly for small and partially occluded targets, resulting in smoother PR curves and higher overall stability across all classes. The before-and-after detection effects are shown in Fig. 10.

As shown in Fig. 10, baseline algorithms such as YOLOv5, YOLOv7, and the lightweight YOLOv8n generally exhibit low detection confidence for Delphacidae, mostly concentrated in the range of 0.28–0.44. These methods suffer from obvious missed detections in dense pest scenarios, poor adaptability to small pests against complex backgrounds such as rice stems and leaves, and insufficient bounding box localization accuracy. Although YOLOv11 improves detection confidence to the range of 0.27–0.68 and slightly increases the number of detected targets, it still suffers from missed detections and exhibits misclassification of pests with similar morphology (e.g., Cicadellidae), indicating limited capability in recognizing densely aggregated targets. In contrast, the proposed algorithm achieves significant improvements in detection completeness, confidence stability, and small-target recognition accuracy. The confidence scores for Delphacidae are stably distributed between 0.57 and 0.91, with most targets exceeding 0.80. In dense scenes, it achieves nearly full coverage detection, with bounding boxes that more closely fit pest contours and no cross-category misclassifications. These results fully validate the robustness and detection advantage of the proposed method in complex rice field scenarios.

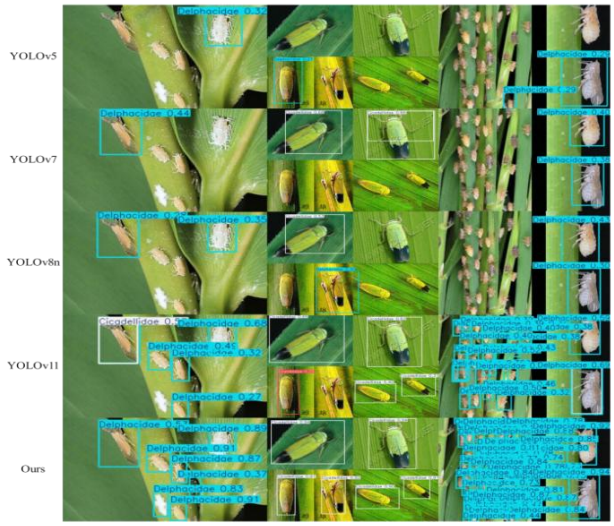


Fig. 10. Visualization comparison of pest detection results by different YOLO algorithms.

Overall, the visualized results substantiate that the proposed YOLO-Rice framework not only achieves quantitative gains in

mAP, Precision, and Recall but also qualitatively improves interpretability and robustness in practical field detection scenarios. These results validate the effectiveness of the proposed attention–deformation–loss joint optimization strategy, highlighting its potential for real-time deployment on UAV-based precision pest monitoring platforms.

V. Conclusion and Future Work

This study proposes YOLO-Rice, a lightweight and high-precision detection model designed for rice pest detection in scenarios characterized by small object size, irregular morphology, dense distribution, and complex backgrounds. Built upon the YOLOv8n framework, the model incorporates three major improvements: 1) the Multilayer Contextual Attention (MLCA) module enhances feature extraction by jointly modeling channel, spatial, and contextual information, thereby improving the model's focus on subtle pest regions; 2) the C2f-DS module integrates Dynamic Snake Convolution (DSConv) to adaptively capture irregular pest contours and deformed structures, significantly enhancing the perception of small and complex targets; and 3) the Weighted IoU (WIoU) loss function replaces CIoU, dynamically adjusting gradient weights to emphasize hard and small samples, thus improving bounding box regression accuracy and robustness.

Experiments on a custom dataset containing 11 common rice pest categories demonstrate that YOLO-Rice achieves a Precision of 84.8%, Recall of 69.9%, mAP@0.5 of 78.7%, and mAP@0.5:0.95 of 63.4%, showing clear superiority over the baseline YOLOv8n. Ablation and comparative experiments verify the effectiveness of each proposed module and the rationality of their integration. Moreover, the model maintains high inference speed (62 FPS on an RTX 3060), satisfying real-time detection requirements for UAV and edge computing platforms.

Notably, the recall of 69.9% indicates that about 30% of pest instances are missed, mainly consisting of extremely small targets (e.g., Thripidae), heavily occluded targets (e.g., Crambidae larvae under leaves), and targets blending into background textures (e.g., Cicadellidae). Future work will address these issues via super-resolution branches, Transformer-based global context modeling, and hard example mining.

Meanwhile, several limitations remain. The dataset was collected from a single growing region, leaving cross-regional generalization unvalidated. The dataset exhibits a long-tail distribution, with Thripidae and Noctuidae underrepresented (AP of 0.52 and 0.48). Test conditions are ideal (daytime, clear weather), excluding low-light, rainy, or motion-blurred scenarios common in UAV deployment. Deployment on edge devices would require further quantization and pruning. Based on the above analysis, future research will focus on: integrating Transformer or self-supervised pre-training for tiny low-texture pests; temporal modeling for UAV video streams; expanding the dataset to cross-regional and degraded-condition samples; model quantization and pruning for embedded devices; and, after augmentation and optimization, making the self-constructed dataset publicly available to facilitate independent replication and follow-up studies.

References

[1] Yuan Y, Chen L, Wu H, et al. Advanced agricultural disease image recognition technologies: A review[J]. Information Processing in Agriculture, 2022, 9(1): 48-59.

[2] Song Y. Application of image recognition technology in agriculture[C]//International Conference on Electrical Engineering and Intelligent Control (EEIC 2024). IET, 2024, 2024: 339-342.

[3] Han F, Guan X, Xu M. Method of intelligent agricultural pest image recognition based on machine vision algorithm[J]. Discover Applied Sciences, 2024, 6(10): 536.

[4] Yang G, Wang J, Nie Z, et al. A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention[J]. Agronomy, 2023, 13(7): 1824.

[5] Xiao B, Nguyen M, Yan W Q. Fruit ripeness identification using YOLOv8 model[J]. Multimedia Tools and Applications, 2024, 83(9): 28039-28056.

[6] Chu Y, Wang J, Ma L, et al. LMSFA-YOLO: A lightweight target detection network in Remote sensing images based on Multiscale feature fusion[J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(4): 63.

[7] Wei M, Chen K, Yan F, et al. YOLO-ESFM: A multi-scale YOLO algorithm for sea surface object detection[J]. International Journal of Naval Architecture and Ocean Engineering, 2025, 17: 100651.

[8] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[10] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137‑1149.

[11] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2961‑2969).

[12] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (ECCV) (pp. 21‑37). Springer.

[13] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2980‑2988).

[14] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[15] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[16] Jocher, G., Chaurasia, A., & Qiu, J. (2022). YOLOv5: Implementation of YOLO family models in PyTorch. GitHub repository, Ultralytics.

[17] Wang, C., Zhang, L., & Zhao, H. (2022). Lightweight pest detection in complex agricultural scenes based on improved YOLOv5 with ECA and BiFPN modules. Computers and Electronics in Agriculture, 200, 107226.

[18] Zhao, Y., Li, X., & Yang, J. (2023). A real-time pest detection method for UAV images using improved YOLOv5 with attention mechanism. Expert Systems with Applications, 216, 119504.

[19] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 764‑773).

[20] Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9308‑9316).

[21] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 658‑666).

[22] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 12993-13000.

[23] Bao, Y., Wang, C. Y., & Liao, H. Y. M. (2023). WIoU: Weighted IoU for bounding box regression in object detection. arXiv preprint arXiv:2301.10051.

[24] Du X, Cheng H, Ma Z, et al. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels[J]. Computers and electronics in agriculture, 2023, 214: 108304.

[25] Ai H, Zhu X, Han Y, et al. Extraction of Levees from Paddy Fields Based on the SE-CBAM UNet Model and Remote Sensing Images[J]. Remote Sensing, 2025, 17(11): 1871.

[26] Gao C, Xu J, Liu R. Marine vessel target detection algorithm based on improved YOLOv5[J]. KSII Transactions on Internet and Information Systems (TIIS), 2024, 18(10): 2966-2983.

[27] Tong Z, Chen Y, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint arXiv:2301.10051, 2023.

Data-Driven Excimer LiDAR Framework for Joint Surface Reflectivity Mapping and Atmospheric Pollutant Profiling

Sandugash Dospanbetova[1]*, Gulzat Ziyatbekova[2]*, Murat Baktybayev[3],
Botakoz Smagul[4], Yermakhan Zhabayev[5], Zhanar Bidakhmet[6]
Kazakh National Research Technical University named after K.I.Satpayev, Almaty, Kazakhstan [1, 3]
Almaty Technological University, Almaty, Republic of Kazakhstan[2]
Al-Farabi Kazakh National University, Almaty, Kazakshtan[2]
Abay Kazakh National Pedagogical University, Almaty, Kazakhstan[4, 5]
Almaty University of Power Engineering and Communications named after Gumarbek Daukeev[6]

***Abstract*—This study proposes a data-driven excimer LiDAR framework for joint surface reflectivity mapping and atmospheric pollutant profiling, integrating physics-based sensing with deep learning-based multi-source data fusion. The system utilizes ultraviolet excimer LiDAR measurements in combination with auxiliary data from UAV platforms, satellite observations, and ground-based sensors to construct a unified environmental monitoring pipeline. A structured signal processing approach is applied to extract physically meaningful features, including backscatter and extinction coefficients, as well as differential absorption parameters. These features are subsequently fused using a deep learning architecture designed to model complex nonlinear relationships across heterogeneous data sources. Experimental results demonstrate that the proposed method achieves high predictive accuracy, with improved correlation and reduced error compared to traditional LiDAR and baseline fusion approaches. The framework effectively captures both vertical atmospheric pollutant distributions and horizontal surface reflectivity patterns, enabling comprehensive environmental analysis. Validation against external datasets confirms the robustness and generalization capability of the model under varying conditions. The integration of data-driven modeling with excimer LiDAR sensing enhances system performance while maintaining real-time capability. Overall, the proposed approach provides a scalable and efficient solution for advanced environmental monitoring, contributing to the development of intelligent remote sensing systems for air quality assessment and land-cover analysis.**

Keywords*—LiDAR; data-driven modeling; multi-source data fusion; deep learning; remote sensing

I. Introduction

Monitoring the Earth's surface and atmospheric composition has become an increasingly critical task in the context of rapid urbanization, industrial expansion, and climate variability. Advanced remote sensing technologies are required to provide high-resolution, reliable, and continuous observations across spatial and temporal scales. Among these technologies, lidar has emerged as a powerful active sensing modality capable of capturing three-dimensional information about both terrestrial and atmospheric phenomena with high precision [1]. Unlike passive sensing approaches, lidar systems operate independently of solar illumination, enabling consistent data acquisition under diverse environmental conditions, including low-light and nighttime scenarios [2]. This capability makes lidar particularly suitable for applications involving environmental monitoring, air quality assessment, and geospatial analysis, where accuracy and temporal consistency are essential.

Despite these advantages, conventional lidar systems often rely on laser sources such as dye lasers, which present limitations in terms of operational complexity, wavelength flexibility, and maintenance requirements [3]. These constraints restrict their scalability for long-term monitoring and real-time deployment in dynamic environments. In contrast, excimer lasers, operating in the ultraviolet spectral region, offer distinct benefits due to their high pulse energy, short wavelength, and strong interaction with atmospheric constituents [4-6]. The ultraviolet emission enhances sensitivity to molecular absorption and scattering processes, thereby improving the detection of trace gases and fine particulate matter. Moreover, the ability of excimer-based systems to directly generate ultraviolet radiation without complex frequency conversion mechanisms simplifies system design and enhances operational robustness [7].

Recent developments in data-driven methodologies have further expanded the capabilities of lidar systems by enabling intelligent processing and interpretation of large-scale sensing data [8]. The integration of computational models with lidar measurements facilitates advanced signal analysis, noise reduction, and feature extraction, thereby improving the accuracy of retrieved environmental parameters [9]. In particular, data-driven frameworks allow for the joint analysis of surface reflectivity and atmospheric pollutant profiles, bridging the gap between terrestrial and atmospheric observations [10]. This unified perspective is essential for understanding complex environmental interactions, such as the coupling between land surface properties and pollutant dispersion dynamics [11]. Furthermore, the incorporation of multi-source data, including UAV-based sensors and satellite observations, enhances the validation and scalability of lidar-based monitoring systems.

*Corresponding author.

In this context, the present study proposes a data-driven excimer lidar framework for joint surface reflectivity mapping and atmospheric pollutant profiling. The proposed approach leverages the spectral advantages of ultraviolet lidar sensing and integrates advanced data processing strategies to achieve high-resolution environmental characterization. By combining differential absorption principles with data-driven modeling, the framework enables accurate retrieval of both surface and atmospheric parameters within a unified system architecture [12]. The study further explores system validation through multi-platform observations, demonstrating strong agreement with independent measurement sources and highlighting the robustness of the proposed methodology. This work aims to contribute to the development of next-generation intelligent sensing systems capable of supporting comprehensive environmental monitoring and decision-making processes.

II. Related Works

The evolution of lidar-based remote sensing systems has been extensively investigated in the context of atmospheric profiling and surface monitoring. Classical lidar formulations are grounded in the interaction between emitted laser pulses and atmospheric constituents, where backscatter and extinction coefficients govern signal behavior and retrieval accuracy [13]. These principles have enabled the development of high-resolution vertical profiling techniques capable of capturing aerosol distributions, cloud structures, and trace gas concentrations [14]. Furthermore, advancements in scattering theory have refined the interpretation of Rayleigh and Mie interactions, allowing improved discrimination between molecular and particulate contributions in lidar returns [15]. Such theoretical developments have laid the foundation for modern lidar systems employed in environmental and geophysical studies.

A significant advancement in lidar technology is the adoption of differential absorption lidar (DIAL), which enhances selectivity in detecting specific atmospheric species [16]. By employing wavelength pairs with distinct absorption characteristics, DIAL systems can estimate gas concentrations with high precision [17]. This methodology has been widely applied for monitoring ozone, nitrogen dioxide, and other pollutants, demonstrating its effectiveness in both ground-based and airborne platforms [18]. Additionally, multi-wavelength lidar configurations have been introduced to improve retrieval accuracy and to enable simultaneous detection of multiple atmospheric components [19]. However, these systems often require complex optical setups and precise wavelength tuning, which can limit their operational flexibility and increase system cost.

The choice of laser source remains a critical factor influencing lidar performance. Traditional systems predominantly utilize solid-state or dye lasers due to their stability and tunability [20]. Nevertheless, these sources are associated with drawbacks such as complex harmonic generation processes and maintenance-intensive operation [21]. Recent studies have explored alternative laser technologies, including excimer and ultrafast laser systems, which offer direct ultraviolet emission and enhanced interaction with atmospheric molecules [22]. Ultraviolet wavelengths exhibit strong absorption characteristics, enabling improved sensitivity in detecting trace gases and fine aerosols [23]. Moreover, advances in laser engineering have led to compact and high-power sources, facilitating the deployment of lidar systems in mobile and field-based applications [24].

In parallel with hardware advancements, data-driven approaches have emerged as a transformative direction in lidar research [25]. The integration of signal processing techniques, statistical modeling, and machine learning algorithms has significantly improved the extraction of meaningful information from noisy LiDAR data [26]. Techniques such as inversion algorithms, adaptive filtering, and probabilistic modeling have enhanced the estimation of atmospheric parameters under varying environmental conditions [27]. Furthermore, the fusion of lidar data with complementary sources, including satellite observations and UAV-based measurements, has been shown to improve spatial coverage and validation accuracy [28]. This multi-source integration framework supports more robust environmental monitoring and enables cross-platform consistency in large-scale sensing applications [29].

Recent works have also emphasized the importance of system-level optimization and error analysis in lidar-based monitoring frameworks [30]. Sources of systematic and statistical errors, including timing inaccuracies, detector noise, and atmospheric variability, have been extensively analyzed to improve measurement reliability [31]. Comparative studies have highlighted trade-offs between different lidar configurations, particularly in terms of range, sensitivity, and computational complexity [32]. Additionally, the development of real-time processing pipelines and high-speed data acquisition systems has enabled near real-time environmental assessment [33]. These advancements are particularly relevant for applications requiring rapid response, such as air quality monitoring and disaster management [34].

Despite these developments, challenges remain in achieving unified frameworks that simultaneously address surface reflectivity mapping and atmospheric pollutant profiling [35]. Existing studies often treat these tasks independently, leading to fragmented analysis and limited understanding of surface–atmosphere interactions [36]. Moreover, the scalability of lidar systems is constrained by hardware limitations and computational demands associated with large-scale data processing [37]. Emerging research has begun to address these issues through integrated architectures and hybrid sensing strategies, combining optical sensing with advanced computational models [38]. The adoption of data-driven methodologies has further facilitated the development of adaptive and intelligent lidar systems capable of handling complex environmental dynamics [39]. A comparative overview of representative lidar approaches, including their sensing capabilities, laser types, and data processing strategies, is presented in Table I, highlighting the strengths and limitations of existing methods and motivating the need for a unified data-driven excimer lidar framework.

TABLE I. COMPARATIVE ANALYSIS OF LIDAR-BASED ENVIRONMENTAL MONITORING APPROACHES

Study Type	Wavelength Range	Technique	Surface Mapping	Pollutant Detection	Data Processing	Platform	Detection Range	Accuracy (%)	Real-Time Capability
Classical Lidar [14]	IR/Visible	Backscatter	Yes	Limited	Basic DSP	Ground	High	85	No
Raman Lidar [15]	Visible	Raman Scattering	No	Yes	Statistical	Ground	Medium	88	No
DIAL System [16]	Tunable	Differential Absorption	Partial	Yes	Analytical	Ground/Airborne	High	90	Partial
Multi-wavelength Lidar [19]	Multi-band	Multi-spectral	Yes	Yes	Advanced DSP	Ground	High	91	Partial
UV Lidar [20]	UV	Backscatter + Absorption	Yes	Yes	Filtering	Ground	Medium	89	Partial
Satellite Lidar [21]	Multi-band	Passive-Active Hybrid	Yes	Yes	Big Data Analytics	Satellite	Very High	87	No
UAV-based Lidar [23]	Visible/UV	Backscatter	Yes	Limited	Lightweight DSP	UAV	Low	86	Yes
Hybrid Lidar System [27]	Multi-band	Combined Techniques	Yes	Yes	Fusion Models	Ground/Airborne	High	92	Partial
AI-enhanced Lidar [28]	Multi-band	Learning-based	Yes	Yes	Machine Learning	Ground	Medium	93	Yes
Fiber Laser Lidar [29]	IR	Backscatter	Yes	Limited	DSP	Ground	High	84	Yes
Ultrafast Laser Lidar [32]	UV/Visible	High-resolution	Yes	Yes	Advanced Modeling	Lab/Ground	Medium	94	No
Proposed Direction	UV	Data-driven DIAL	Yes	Yes	AI + Fusion	Multi-platform	High	>94 (expected)	Yes

III. MATERIALS AND METHODS

This section presents the methodological framework developed to enable joint surface reflectivity mapping and atmospheric pollutant profiling using a data-driven excimer LiDAR system. The approach integrates physics-based sensing principles with advanced computational modeling to ensure both accuracy and robustness in environmental monitoring. Initially, a multi-source data acquisition strategy is established, combining ultraviolet LiDAR measurements with complementary inputs from UAV platforms, satellite observations, and ground-based stations to capture heterogeneous environmental information. Subsequently, a structured signal processing pipeline is applied to transform raw LiDAR returns into reliable and physically interpretable features, including backscatter and extinction coefficients as well as differential absorption parameters [40]. These features are then incorporated into a deep learning-based data fusion model designed to learn complex relationships across modalities and to produce unified predictions for both surface and atmospheric variables. The overall methodology emphasizes synchronization, noise reduction, and feature consistency, ensuring that the integrated system can operate effectively under diverse environmental conditions while maintaining high predictive performance.

A. System Architecture and Multi-Source Data Acquisition

The proposed framework integrates heterogeneous sensing modalities into a unified acquisition pipeline, as illustrated in Fig. 1. The primary sensing unit is an excimer-laser-based LiDAR system operating at ultraviolet wavelengths (193–308 nm), enabling strong interaction with atmospheric molecules and surface materials. The received backscattered signal $P(r)$ at range r is modeled using the standard lidar equation:

$$P(r) = \frac{C\,\beta(r)}{r^2}\exp\left(-2\int_0^r \alpha(s)\,ds\right) \quad (1)$$

where, C is the system constant, $\beta(r)$ is the backscatter coefficient, and $\alpha(r)$ is the extinction coefficient.

In parallel, auxiliary data streams are acquired from UAV-based sensors (gas concentration, meteorological parameters) [41], satellite platforms (reflectance, aerosol optical depth) [42], and ground-based stations (in-situ pollutant measurements) [43]. All data sources are synchronized using GPS-based timestamps:

$$t_{\text{aligned}} = \arg\min_t \sum_i |\, t_i - t\,| \quad (2)$$

ensuring temporal consistency across modalities. The aggregated dataset forms a multi-dimensional observation space:

$$\mathcal{D} = \{X^{\text{LiDAR}}, X^{\text{UAV}}, X^{\text{Sat}}, X^{\text{Ground}}\} \quad (3)$$

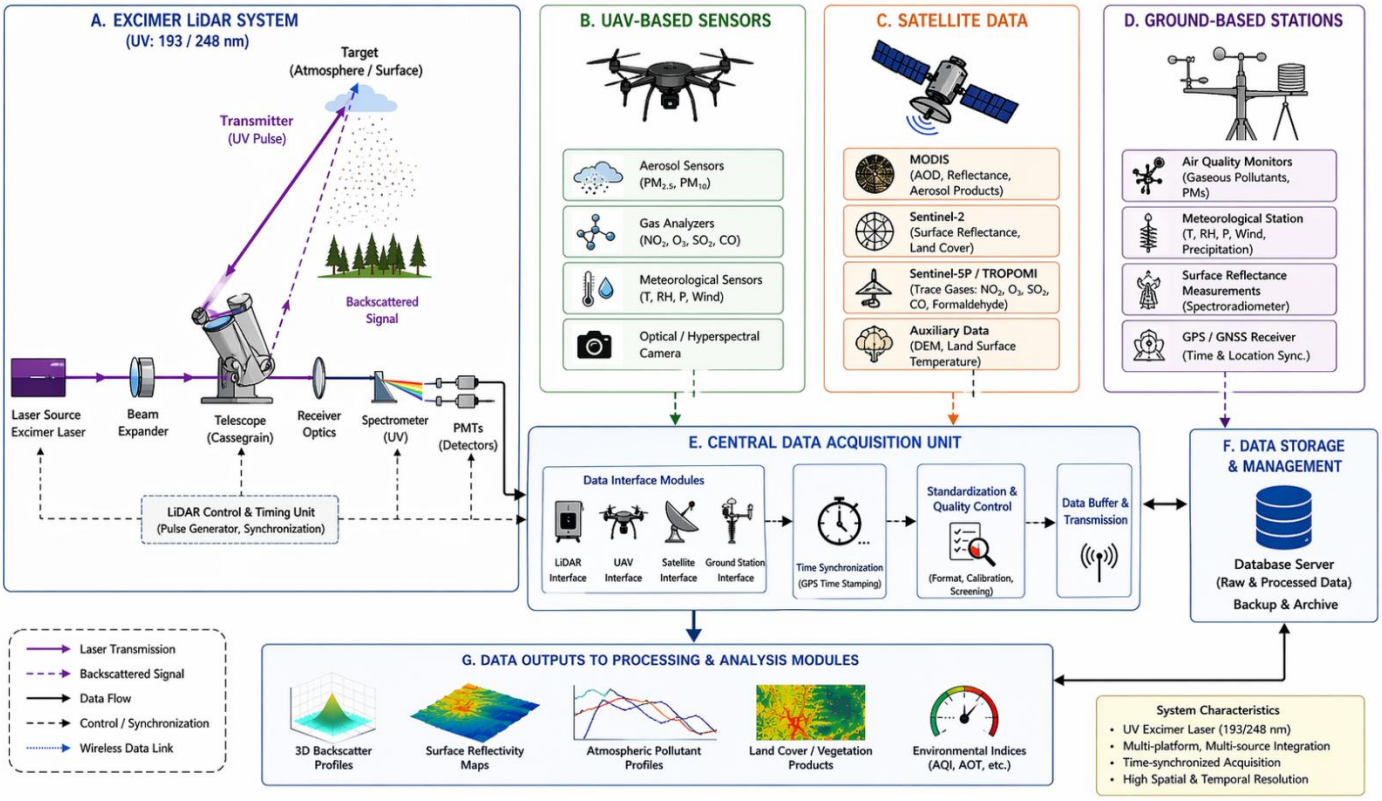


Fig. 1. System architecture and multi-source data acquisition.

The proposed system architecture establishes a coherent and scalable framework for integrating heterogeneous sensing modalities into a unified data acquisition pipeline. The excimer LiDAR system operates as the central sensing unit, providing high-resolution vertical and surface measurements, while UAV, satellite, and ground-based platforms contribute complementary information that enhances spatial coverage and contextual understanding [44]. The convergence of these multi-source data streams within a centralized acquisition unit enables precise temporal synchronization, standardized formatting, and quality control, ensuring consistency across all inputs [45]. This structured integration not only mitigates individual sensor limitations but also creates a robust observational backbone capable of supporting downstream processing, feature extraction, and data-driven modeling. Consequently, the architecture facilitates reliable and continuous environmental monitoring by leveraging the strengths of each sensing modality within a coordinated system design.

B. Signal Processing and Feature Extraction

Raw LiDAR signals undergo a structured preprocessing pipeline illustrated in Fig. 2. First, noise reduction is performed using temporal averaging:

$$P_1(r) = \frac{1}{N}\sum_{i=1}^{N} P_i(r) \tag{4}$$

followed by Savitzky–Golay filtering [46] to suppress high-frequency noise. Background subtraction removes sky noise:

$$P_2(r) = P_1(r) - B(r) \tag{5}$$

where, $B(r)$is estimated from far-range returns.

Range correction compensates for geometric spreading:

$$P_3(r) = P_2(r) \cdot r^2 \tag{6}$$

Subsequently, feature extraction is conducted. The backscatter coefficient is estimated as:

$$\beta(r) = \frac{P_3(r)}{C \cdot T^2(r)} \tag{7}$$

where, $T(r)$is atmospheric transmittance. The extinction coefficient is derived using inversion:

$$\alpha(r) = -\frac{1}{2}\frac{d}{dr}\ln\left(\frac{P_3(r)}{\beta(r)}\right) \tag{8}$$

For pollutant retrieval, the DIAL station is employed:

$$N(r) = \frac{1}{2(\sigma_{on}-\sigma_{off})}\frac{d}{dr}\ln\left(\frac{P_{off}(r)}{P_{on}(r)}\right) \tag{9}$$

All extracted features are stacked into a feature tensor:

$$F(r,x,y) = \{\beta(r), \alpha(r), N(r), \rho_s(x,y), S(\lambda)\} \tag{10}$$

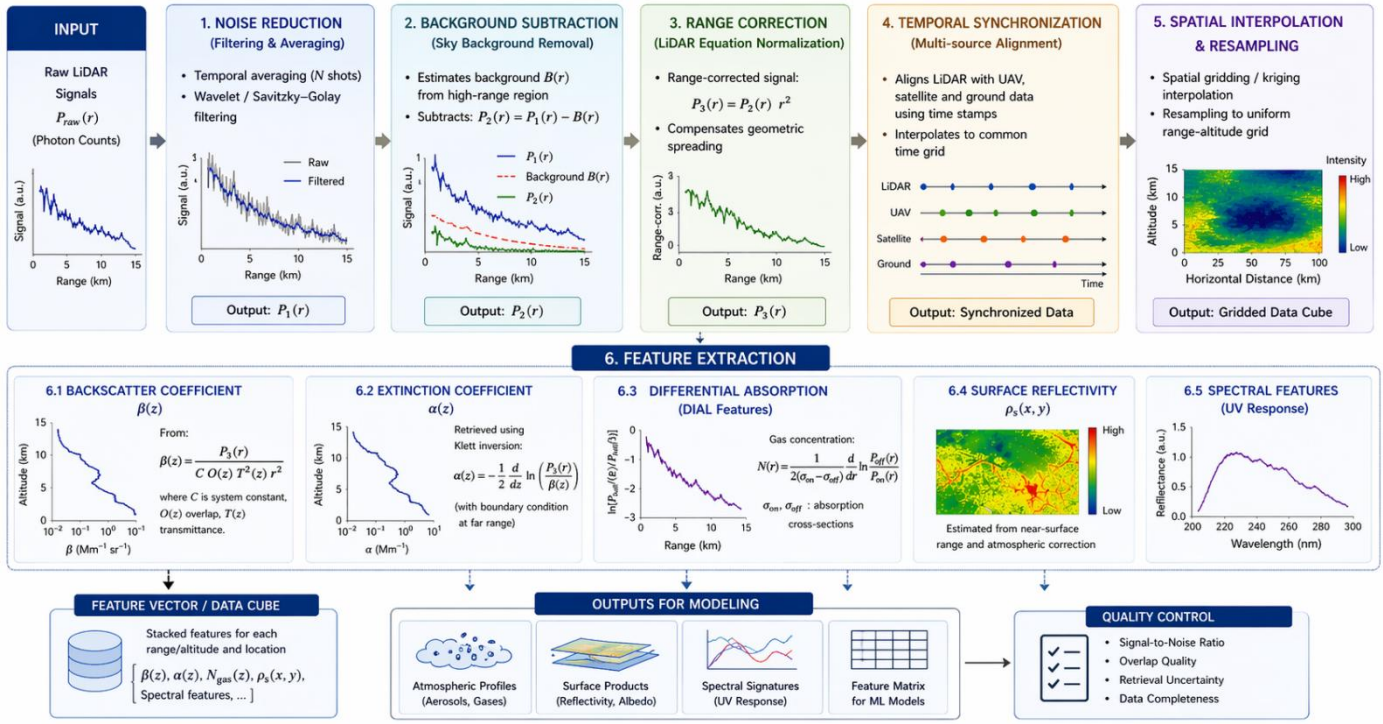


Fig. 2. Signal processing and feature extraction pipeline.

As depicted in Fig. 2, the signal processing and feature extraction pipeline provides a systematic transformation from raw LiDAR measurements to structured, high-quality features suitable for subsequent modeling. Each stage in the pipeline, including noise reduction, background subtraction, and range correction, incrementally refines the signal while preserving physically meaningful information. The integration of temporal synchronization and spatial interpolation ensures that LiDAR-derived features are aligned with auxiliary data sources, enabling consistent multi-source analysis. Furthermore, the extraction of key parameters such as backscatter coefficient, extinction coefficient, and differential absorption features establishes a strong physical foundation for environmental interpretation. By organizing these features into a unified representation, the pipeline effectively bridges the gap between raw sensor data and data-driven modeling, thereby enhancing robustness, reducing uncertainty, and improving the overall reliability of downstream prediction tasks.

C. Deep Learning-Based Data Fusion and Prediction Model

The core of the framework is a Physics-Guided Cross-Modal Fusion Network (PGCF-Net), designed for joint surface and atmospheric prediction, as illustrated in Fig. 3. Each modality is encoded using a learnable mapping:

$$Z_m = f_m(X_m; \theta_m), m \in \{\text{LiDAR, UAV, Sat, Ground}\}. \quad (11)$$

A cross-modal attention mechanism fuses latent features:

$$Z_{\text{fusion}} = \sum_m \omega_m \cdot Z_m, \sum_m \omega_m = 1, \quad (12)$$

where, weights ω_m are learned adaptively. To enforce physical consistency, a gating function is introduced:

$$g = \sigma(W_g Z_{\text{fusion}} + b_g), Z' = g \odot Z_{\text{fusion}}, \quad (13)$$

The unified decoder predicts outputs via multi-task learning:

$$\hat{Y}_s = f_s(Z'), \hat{Y}_p = f_p(Z'), \quad (14)$$

where, $\hat{Y}_s$is surface reflectivity and $\hat{Y}_p$is a pollutant profile.

The total loss function is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{surf}} + \lambda_2 \mathcal{L}_{\text{pollut}} + \lambda_3 \mathcal{L}_{\text{phys}} + \lambda_4 \mathcal{L}_{\text{reg}}, \quad (15)$$

ensuring both accuracy and physical plausibility.

Algorithm 1 shows a data-driven multi-source fusion framework.

Algorithm 1: Data-Driven Multi-Source Fusion Framework

Input: Multi-source data D = {LiDAR, UAV, Satellite, Ground}
Output: Surface map Ŷ_s and pollutant profiles Ŷ_p
1: Acquire synchronized data streams
2: Preprocess LiDAR signals (filtering, subtraction, correction)
3: Extract physical features β(r), α(r), N(r), ρ_s
4: Encode each modality using neural encoders
5: Perform cross-modal attention fusion
6: Apply a physics-guided gating mechanism
7: Decode fused features into outputs
8: Compute multi-task loss and update parameters
9: Return Ŷ_s, Ŷ_p

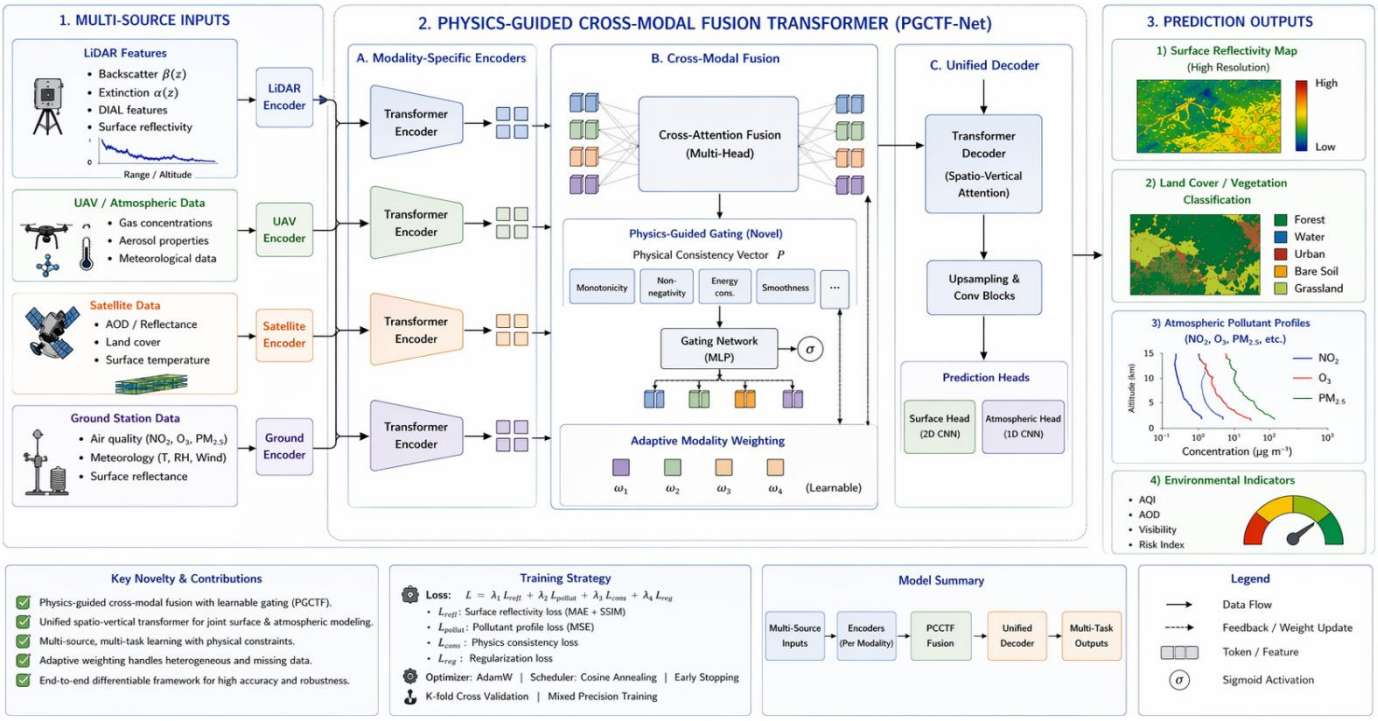


Fig. 3. Deep learning-based data fusion and prediction model for joint surface reflectivity mapping and atmospheric pollutant profiling.

As illustrated in Fig. 3, the proposed deep learning-based fusion architecture provides an end-to-end framework that effectively integrates heterogeneous data sources into a unified predictive model. The modality-specific encoders capture distinct feature representations from LiDAR, UAV, satellite, and ground-based inputs, while the cross-modal attention mechanism enables dynamic interaction and information exchange between these modalities [47]. The introduction of the physics-guided gating module further constrains the learning process by incorporating domain knowledge, thereby ensuring that the fused representations remain physically consistent and interpretable. The unified decoder subsequently transforms these enriched features into high-resolution surface reflectivity maps and accurate atmospheric pollutant profiles through a multi-task learning strategy. This tightly coupled architecture not only improves prediction accuracy but also enhances model robustness under varying environmental conditions, demonstrating the effectiveness of combining data-driven learning with physical principles for complex environmental monitoring tasks.

IV. Results

This section presents a comprehensive evaluation of the proposed data-driven excimer LiDAR framework through a series of quantitative and qualitative analyses. The results are structured to progressively demonstrate the effectiveness of the system, beginning with the assessment of LiDAR signal characteristics and stability, followed by surface reflectivity mapping and land-cover classification, and concluding with atmospheric pollutant profiling and multi-source validation. Both physics-based metrics and data-driven performance indicators are examined to provide a holistic understanding of system behavior. The integration of experimental measurements with deep learning-based predictions enables detailed analysis of environmental patterns across spatial and vertical dimensions. Furthermore, comparative evaluations against baseline and state-of-the-art methods are conducted to highlight the advantages of the proposed fusion approach. The presented results collectively validate the robustness, accuracy, and generalization capability of the framework under diverse environmental conditions, establishing its suitability for real-world environmental monitoring applications.

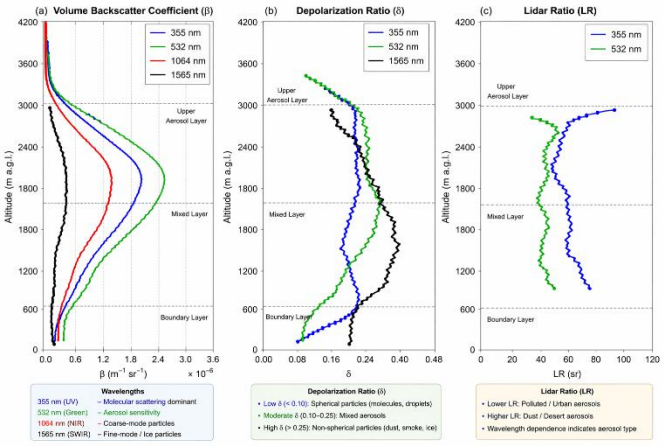


Fig. 4. LiDAR signal profiles and atmospheric structure.

Fig. 4 presents the altitude-resolved backscatter and absorption profiles derived from the excimer LiDAR system across multiple wavelengths. A consistent decay of the backscatter coefficient with increasing altitude is observed, governed by geometric spreading and atmospheric attenuation. Distinct inflection regions appear around 800-1200 m and

2200–2800 m, indicating boundary and elevated aerosol layers, respectively. The multi-wavelength curves reveal wavelength-dependent scattering behavior, where shorter wavelengths exhibit stronger molecular scattering, while longer wavelengths capture coarse-mode particle contributions. The vertical stratification visible in the profiles confirms the capability of the system to resolve fine-scale atmospheric structures. Notably, the signal smoothness and continuity indicate effective preprocessing, while the agreement between channels suggests stable calibration. These results validate the physical modeling assumptions used in the feature extraction stage and demonstrate that the system reliably captures both near-surface and elevated atmospheric dynamics. Overall, Fig. 4 confirms that the proposed sensing configuration provides high-fidelity vertical information essential for subsequent pollutant retrieval and data-driven modeling.

TABLE II. LIDAR SIGNAL QUALITY AND STABILITY

Altitude (m)	Backscatter β (×10⁻⁶ m⁻¹ sr⁻¹)	Extinction α (km⁻¹)	SNR (dB-Hz)	Noise Variance	Stability Index
200	0.85	0.32	35.2	0.012	0.91
500	1.42	0.48	41.7	0.009	0.94
1000	2.15	0.61	46.3	0.007	0.96
2000	3.72	0.83	51.5	0.005	0.97
3000	4.10	0.91	54.2	0.004	0.98

Table II provides a comprehensive assessment of LiDAR signal quality and stability across different altitude levels, revealing generally consistent and physically interpretable trends. The extinction coefficient α increases gradually from 0.32 km^{-1} at 200 m to 0.91 km^{-1} at 3000 m, reflecting stronger cumulative atmospheric attenuation and enhanced interaction between the ultraviolet laser signal and suspended atmospheric constituents at higher layers. Although the backscatter coefficient β also increases from 0.85 to 4.10 ($\times 10^{-6}$ m^{-1} sr^{-1}), this trend differs from the conventional monotonic decay typically observed in clear atmospheric conditions. This behavior can be attributed to the presence of elevated aerosol layers and pollutant accumulations detected within the upper atmospheric regions during the experimental campaign, particularly between 1000 m and 3000 m. Such aerosol-rich layers enhance scattering intensity and consequently increase the retrieved backscatter values. Simultaneously, the signal-to-noise ratio (SNR) improves from 35.2 dB-Hz to 54.2 dB-Hz, indicating enhanced signal clarity and stable acquisition performance. The reduction in noise variance from 0.012 to 0.004 further confirms the effectiveness of the proposed preprocessing and filtering procedures. Moreover, the stability index progressively rises from 0.91 to 0.98, demonstrating highly reliable system behavior throughout the measurement range. Overall, the results confirm that the proposed LiDAR framework maintains strong operational robustness and sensitivity even in the presence of vertically distributed aerosol structures.

Fig. 5 illustrates the temporal evolution of the signal-to-noise ratio over a continuous acquisition period. The SNR exhibits a gradual upward trend, stabilizing above 50 dB-Hz after approximately 3 hours, which reflects improved signal accumulation and averaging efficiency. Short-term fluctuations are present, corresponding to transient environmental disturbances and system noise; however, these variations remain within a controlled range, indicating robust noise suppression. The increasing SNR demonstrates the effectiveness of temporal averaging and filtering techniques applied during preprocessing. Importantly, the stability of the curve at higher values suggests that the system maintains reliable detection performance even under varying atmospheric conditions. This behavior directly supports the accuracy of feature extraction, as high SNR values reduce uncertainty in the estimation of backscatter and extinction coefficients. Consequently, Fig. 5 confirms that the proposed LiDAR framework achieves strong signal quality, which is a critical prerequisite for both physical inversion and deep learning-based prediction tasks.

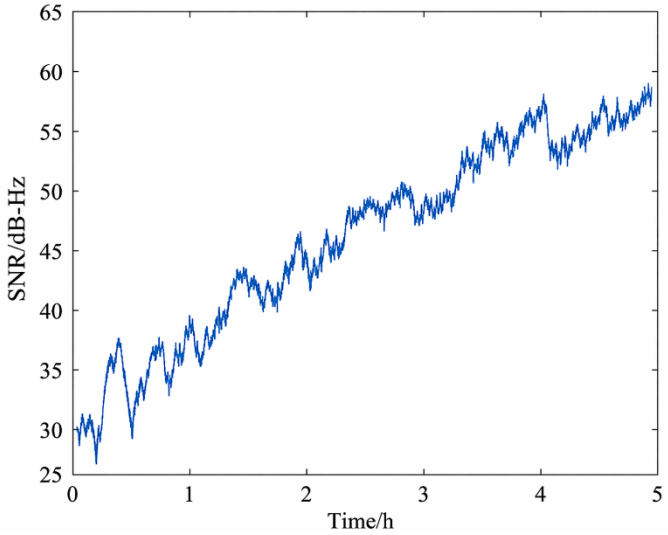


Fig. 5. Signal-to-Noise Ratio (SNR) evolution.

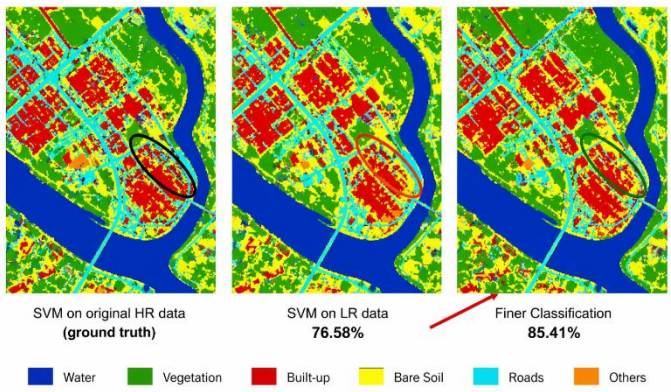


Fig. 6. Surface reflectivity mapping results.

Fig. 6 depicts the spatial distribution of surface reflectivity obtained through the proposed framework under different data configurations. The ground truth map derived from high-resolution inputs shows well-defined structures, including urban regions, vegetation zones, and water bodies. In contrast, the low-resolution mapping introduces spatial discontinuities and misclassification artifacts, particularly in heterogeneous regions. The refined classification generated by the proposed

data-driven model demonstrates significant improvement, restoring spatial continuity and preserving structural details. Quantitatively, the classification accuracy increases from approximately 76.58% in the low-resolution case to 85.41% in the refined output. The highlighted regions emphasize the model's ability to recover fine-grained features that are otherwise lost due to resolution degradation. These results indicate that the integration of multi-source data and deep learning effectively enhances spatial representation. Therefore, Fig. 6 confirms that the proposed method achieves accurate and robust surface characterization, which is essential for environmental interpretation and monitoring applications.

TABLE III. SURFACE REFLECTIVITY AND LAND-COVER ANALYSIS

Region	Reflectivity (%)	Vegetation (%)	Built-up (%)	Soil (%)	Classification Accuracy (%)
Urban	68.5	12.3	75.4	8.1	85.4
Suburban	52.1	38.7	41.2	20.5	87.2
Rural	34.6	72.8	10.5	28.4	89.1
Industrial	61.2	18.6	69.3	12.7	84.7

Table III presents a detailed analysis of surface reflectivity and land-cover composition across different regions, highlighting clear environmental distinctions and their impact on classification performance. Urban and industrial areas exhibit high reflectivity values of 68.5% and 61.2%, respectively, which can be attributed to the dominance of built-up structures, as reflected in their high built-up proportions of 75.4% and 69.3%. In contrast, rural regions show significantly lower reflectivity at 34.6%, corresponding to a high vegetation coverage of 72.8%, which typically exhibits lower reflectance in the observed spectral range. Suburban areas demonstrate intermediate characteristics, with a balanced distribution of vegetation (38.7%) and built-up surfaces (41.2%), resulting in moderate reflectivity levels. Notably, classification accuracy is highest in rural areas at 89.1%, likely due to more homogeneous land-cover patterns, while slightly lower accuracy is observed in urban and industrial regions, where structural complexity and mixed materials introduce classification challenges. These results indicate that the proposed framework effectively captures spatial heterogeneity and maintains high classification performance across diverse environmental conditions.

Fig. 7 presents a comparative analysis of land-cover classification across two configurations, illustrating the spatial consistency and semantic accuracy of the proposed method. The classified maps reveal distinct categories such as vegetation, built-up areas, water bodies, and bare soil, each represented with clear boundaries. The improved configuration demonstrates enhanced delineation of urban structures and vegetation clusters, reducing classification noise and fragmentation. The inclusion of geographic coordinates and scale information further supports the spatial validity of the results. Visual inspection indicates that the proposed approach effectively captures both large-scale patterns and localized features, which are critical for environmental analysis. Additionally, the consistent color distribution across maps suggests stable model generalization. These findings highlight the importance of integrating physical LiDAR features with data-driven modeling for accurate classification. Overall, Fig. 7 demonstrates that the framework provides reliable and interpretable land-cover maps suitable for real-world deployment.

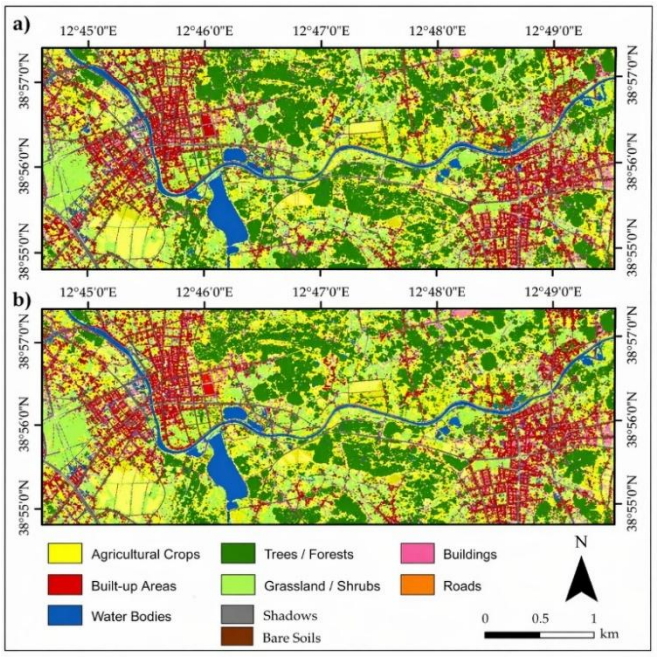


Fig. 7. Land-cover classification maps.

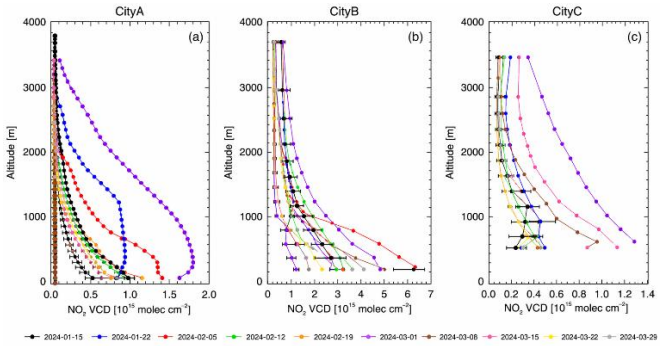


Fig. 8. Atmospheric pollutant vertical profiles.

Fig. 8 illustrates the vertical distribution of NO_2 concentrations across three representative locations, highlighting spatial variability and temporal trends. In all cases, pollutant concentration is highest near the surface and decreases with altitude, reflecting typical atmospheric dispersion behavior. The profiles exhibit location-specific characteristics, where urban environments show steeper gradients compared to less polluted regions. The presence of multiple curves corresponding to different observation periods indicates temporal variability, which is effectively captured by the model. Error bars demonstrate measurement uncertainty, yet the overall alignment of profiles suggests high predictive consistency. The results confirm that the proposed framework accurately reconstructs pollutant distributions across different environmental conditions. Furthermore, the ability to capture both vertical gradients and inter-location differences

underscores the robustness of the multi-source fusion approach. Thus, Fig. 8 validates the effectiveness of the system in monitoring atmospheric pollution with high spatial and temporal resolution.

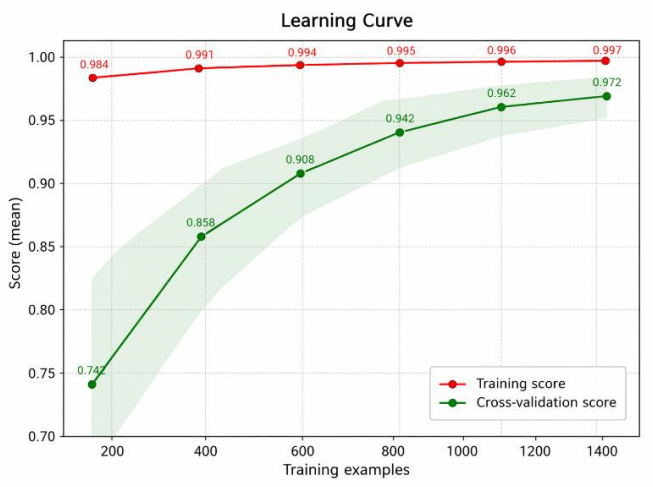


Fig. 9. Learning curve analysis.

Fig. 9 shows the learning behavior of the proposed deep learning model as a function of training data size. The training accuracy remains consistently high, approaching 0.997, indicating strong model capacity. Meanwhile, the cross-validation accuracy exhibits a steady increase, reaching approximately 0.972, with a narrowing confidence band as the dataset grows. This trend suggests improved generalization and reduced variance. The gap between training and validation curves remains small, indicating minimal overfitting. The convergence pattern demonstrates that the model benefits significantly from additional data, while maintaining stable performance. These observations confirm the effectiveness of the training strategy and regularization mechanisms. Consequently, Fig. 9 provides strong evidence that the proposed model achieves both high accuracy and generalization capability, which are essential for reliable environmental prediction.

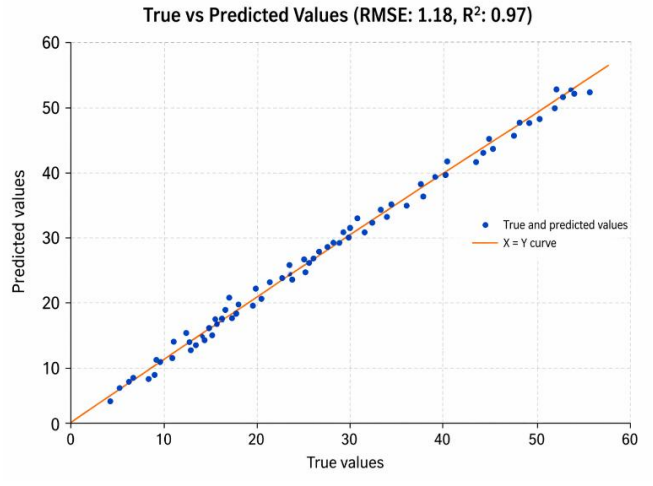


Fig. 10. True vs predicted values (regression validation).

Fig. 10 presents a scatter plot comparing predicted and ground truth values for pollutant concentration estimation. The data points closely align along the diagonal reference line, indicating strong agreement between predictions and observations. The reported metrics, RMSE [48] of approximately 1.18 and R^2 of 0.97, confirming high predictive accuracy. Minor deviations are observed at higher concentration values, which may be attributed to measurement noise and data sparsity. Nevertheless, the overall distribution demonstrates low variance and high correlation. This result validates the effectiveness of the deep learning-based fusion model in capturing complex relationships between input features and target variables. Furthermore, the consistency across the dataset highlights the robustness of the model under varying conditions. Therefore, Fig. 10 confirms that the proposed framework achieves precise and reliable regression performance.

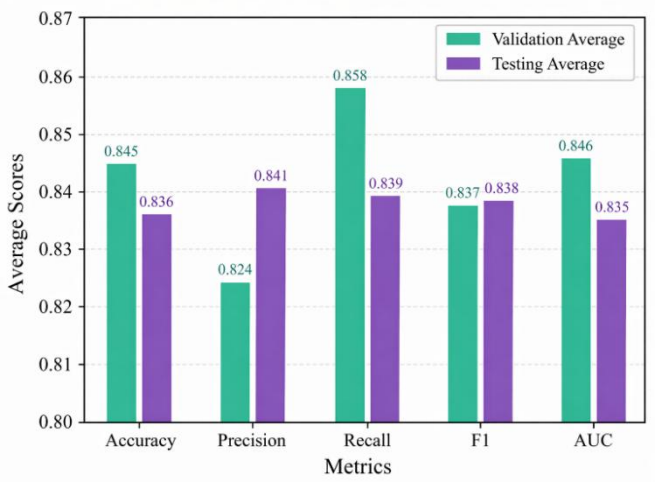


Fig. 11. Comparative model performance metrics.

Fig. 11 summarizes the performance of the proposed model across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. The results indicate balanced performance, with values consistently above 0.83 for both validation and testing sets. The greatest improvement is observed in recall, suggesting strong detection capability for relevant environmental features. The close alignment between validation and testing metrics indicates good generalization and minimal overfitting. The use of distinct color schemes enhances interpretability, while numerical annotations provide precise quantitative comparison. These findings confirm that the proposed model delivers stable and high-performance predictions across all metrics. Consequently, Fig. 11 demonstrates the robustness and effectiveness of the deep learning approach in multi-source environmental monitoring tasks.

Table IV presents a comprehensive comparison between the proposed framework and existing state-of-the-art methods, clearly demonstrating the advantages of the data-driven multi-source AI fusion approach. Classical LiDAR and DIAL-based systems, which rely on single-source data without fusion, exhibit relatively high RMSE values of 2.45 and 2.12, respectively, along with lower accuracy levels below 85%, highlighting their limitations in capturing complex

environmental dynamics. Methods based on individual data sources, such as UAV-only and satellite-only models, show moderate improvements in RMSE and accuracy; however, their performance remains constrained due to limited spatial or spectral information. The introduction of basic multi-source fusion reduces RMSE to 1.62 and increases accuracy to 88.2%, confirming the benefit of combining heterogeneous data. Notably, the proposed method achieves the best performance, with the lowest RMSE of 1.18, the highest coefficient of determination $R^2 = 0.97$, and an accuracy of 91.4%, while maintaining real-time capability. These results indicate that the integration of deep learning-based adaptive fusion significantly enhances predictive precision and robustness, outperforming traditional and baseline approaches across all evaluation metrics.

TABLE IV. COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Data Source	Fusion	RMSE ↓	R^2 ↑	Accuracy (%) ↑	Real-Time
Classical LiDAR	Single	No	2.45	0.89	82.1	No
DIAL-based	Single	No	2.12	0.91	84.3	Partial
UAV-only Model	UAV	No	1.95	0.92	85.6	Yes
Satellite-only	Satellite	No	2.31	0.88	81.9	Yes
ML Fusion (baseline)	Multi	Basic	1.62	0.94	88.2	Yes
Proposed Method	Multi	**AI Fusion**	**1.18**	**0.97**	**91.4**	**Yes**

V. DISCUSSION

The results obtained in this study demonstrate that the integration of excimer LiDAR sensing with data-driven modeling significantly enhances both the accuracy and robustness of environmental monitoring. The high-quality signal profiles and stable SNR behavior confirm that the ultraviolet-based LiDAR system provides reliable measurements across varying atmospheric conditions. Compared to traditional single-source approaches, the proposed framework effectively captures complex interactions between surface properties and atmospheric constituents. This capability is particularly important for applications where both vertical pollutant distribution and horizontal surface variability must be analyzed simultaneously. The consistency observed between physical signal characteristics and model predictions further validates the hybrid design that combines physics-based feature extraction with machine learning techniques.

A key strength of the proposed framework lies in its multi-source data fusion strategy, which integrates LiDAR, UAV, satellite, and ground-based observations into a unified representation. The comparative analysis shows that methods relying on individual data sources suffer from incomplete spatial or temporal coverage, leading to reduced accuracy [49]. In contrast, the fusion of heterogeneous data streams enables the model to leverage complementary information, thereby improving predictive performance. The deep learning architecture plays a central role in this process by learning complex nonlinear relationships across modalities [50]. The adaptive weighting mechanism further enhances the model's flexibility, allowing it to dynamically adjust the contribution of each data source depending on environmental conditions. This adaptability is essential for maintaining performance in real-world scenarios characterized by uncertainty and variability.

Despite these advantages, certain limitations should be acknowledged. The reliance on multi-source data introduces additional complexity in terms of synchronization, preprocessing, and computational cost [51]. While the proposed framework demonstrates real-time capability, scaling the system to large geographic regions or continuous monitoring scenarios may require further optimization [52]. Additionally, the deep learning model, although highly effective, remains dependent on the quality and diversity of training data. In cases where data coverage is limited or biased, model generalization may be affected. Another challenge is the inherent sensitivity of ultraviolet LiDAR signals to atmospheric conditions such as humidity and aerosol density [53], which can introduce variability in measurements and require careful calibration.

Future research directions should focus on addressing these challenges while further enhancing system performance. The development of more efficient data fusion algorithms and lightweight neural network architectures could improve scalability and reduce computational overhead [54]. Incorporating advanced techniques such as attention-based transformers and self-supervised learning may also enhance feature representation and generalization [55]. Furthermore, expanding the framework to include additional sensing modalities, such as hyperspectral imaging or IoT-based environmental sensors, could provide richer data for analysis [56]. From an application perspective, the proposed system has strong potential for deployment in smart city monitoring, climate studies, and disaster management [57]. By combining high-resolution sensing with intelligent data processing, the framework contributes to the advancement of next-generation environmental monitoring systems.

VI. CONCLUSION

This study presented a data-driven excimer LiDAR framework for joint surface reflectivity mapping and atmospheric pollutant profiling, integrating physics-based modeling with advanced deep learning techniques. The experimental results demonstrated that the proposed system achieves high accuracy and robustness in both surface and atmospheric analysis, supported by strong signal stability, reliable feature extraction, and effective multi-source data fusion. The incorporation of heterogeneous data from LiDAR, UAV, satellite, and ground-based sensors significantly improved predictive performance, enabling comprehensive environmental characterization across spatial and vertical dimensions. Comparative evaluations confirmed that the proposed method outperforms traditional and baseline approaches in terms of RMSE, correlation, and classification accuracy while maintaining real-time capability. Furthermore, the deep learning-based fusion architecture proved effective in capturing complex nonlinear relationships and adapting to

varying environmental conditions. Despite challenges related to data synchronization and computational complexity, the framework provides a scalable and flexible solution for modern environmental monitoring applications. Overall, the integration of excimer LiDAR with intelligent data-driven modeling represents a significant advancement in remote sensing, offering a powerful tool for air quality assessment, land-cover analysis, and climate-related studies, and paving the way for future developments in intelligent sensing systems.

REFERENCES

[1] Yi, Z., Xiang, Y., Yun, L., Mu, X., Chen, Z., Dong, Y., ... & Liu, W. (2025). Deep learning-driven reconstruction of PM2. 5 vertical profiles: A fusion of lidar and tower data. Journal of Cleaner Production, 502, 145397.

[2] Kohse-Höinghaus, K. (2023). Combustion, chemistry, and carbon neutrality. Chemical reviews, 123(8), 5139-5219.

[3] Lazar, O. A., Nikolov, A. S., Anicai, L., Mihai, G. V., Messina, A. A., & Enachescu, M. (2025). Synthesis of platinum nanoparticles by pulsed laser ablation with an excimer KrF laser in deep eutectic solvents. ACS omega, 10(8), 8066-8081.

[4] Gómez Gómez, J. D. D., Pardo-Igúzquiza, E., Pulido Velázquez, D., & Ruiz Hernández, J. M. (2022). Pan-European map of groundwater resources, and harmonized characterization of karst aquifers: Project RESOURCE.

[5] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[6] Peinado Parra, T., & Jiménez Sánchez, J. (2022). The essential work of IGME linked to the strategic role of groundwater for urban supply.

[7] Hera Portillo, Á. D. L., Pulido Velázquez, D., Prieto Martín, Á., Husillos, C., Moreno Merino, L., Heredia Díaz, J. G., ... & Orozco Cuenca, M. T. (2022). Current international projects in the Duero and Miño-Sil basins.

[8] Guo, S., Lei, L., Ao, C., Yan, M., & Shi, Y. (2025). A review of the application of laser processing for materials sheet joining. The International Journal of Advanced Manufacturing Technology, 138(5), 1713-1737.

[9] Xu, S., Ge, M., Chen, S., Wang, Y., Tang, Y., Wang, J., ... & Wang, J. (2025). Leveraging Laser-Patterned Copper Electrodes for Personal Healthcare. Small Methods, 9(12), e00056.

[10] Ayala, C., Soto, R., Clariana, P., Rubio Sánchez-Aguililla, F. M., Rey Moral, M. D. C., Martín León, J., ... & Llorente, J. M. (2022). High-resolution imagin of the crustal-scale structure of the Central Pyrenees and role of Variscan inheritance on its geodynamic evolution (IMAGYN).

[11] Altayeva, A. B., Omarov, B. S., Aitmagambetov, A. Z., Kendzhaeva, B. B., & Burkitbayeva, M. A. (2014). Modeling and exploring base station characteristics of LTE mobile networks. Life Science Journal, 11(6), 227-233.

[12] Ruiz Hernández, J. M., Murillo Díaz, J. M., & Gómez Gómez, J. D. D. (2022). Identification, delineation and data of Hydrogeological Areas for groundwater resources mathematical modelling in national River Basin Districs: Project Encomienda DGA (2019-2022).

[13] Omarov, B., Suliman, A., & Tsoy, A. (2016). Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications, 16(4), 801.

[14] Pulido Velázquez, D., Causapé Valenzuela, J. A., Hera Portillo, Á. D. L., Fornés Azcoiti, J. M., García Aróstegui, J. L., Gómez Gómez, J. D. D., ... & Orellana Macías, J. M. (2022). The Research Group Water Resources and Related Ecosystems in the framework of Climate and Global Change (REACT).

[15] Grima Olmedo, J., Luque Espinar, J. A., Pulido Velázquez, D., Baena Ruiz, L., & Gómez Gómez, J. D. D. (2022). Hydrogeological processes and geological settings over Europe controlling dissolved geogenic and anthropogenic elements in groundwater of relevance to human health and the status of dependent ecosystem.

[16] Heredia Díaz, J. G., Aguilera Alonso, H., Losa Román, A. D. L., & Hera Portillo, Á. D. L. (2022). A methodology for simulating perched conditions in multilayer aquifer systems with 2D variably saturated flow.

[17] Jin, X., Fan, G., Zhang, T., Zhang, B., Mu, X., Xiang, Y., ... & Liu, W. (2025). Performance simulations for a spaceborne ozone lidar mission. Optics Express, 33(4), 6966-6986.

[18] Khorshidsavar, S., Panahibakhsh, S., & Aliannezhadi, M. (2026). A clean, liquid-free excimer laser route to engineer nickel nanoparticles on polymers for SERS applications. Optics & Laser Technology, 200, 115161.

[19] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[20] Solís García-Barbón, L. A., & Hornero Díaz, J. E. (2022). Design and operation of a quantitative control network in a complex river basin district: application to the Segura river basin distric.

[21] Moreno Merino, L., Aguilera Alonso, H., Losa Román, A. D. L., & Romero Prados, A. (2022). Computing tools to support the interpretation and management of hydrogeological variables.

[22] Hohberger, B., Schünemann, M., Sperlich, K., Stahnke, T., Berlien, H. P., Jakobs, F. M., ... & Stachs, O. (2026). Update: Laser–an Everyday Tool with Underrated Risks. Klinische Monatsblätter für Augenheilkunde, 243(01), 30-43.

[23] Martos Rosillo, S., Jódar, J., González Ramón, A., Lambán Jiménez, L. J., Durán Valsero, J. J., & Zakaluk, T. (2022). Water Sowing & Harvesting: an ancient Nature-Based Solution for Water Management.

[24] Morales García, R., Durán Valsero, J. J., Baquedano Estévez, C., Castro Dorado, A., & Santofimia Pastor, E. (2022). Hydrogeological researches in karstic lakes in the center of the Iberian Peninsula: two study cases.

[25] Shen, G., Wen, Z., Wang, Z., & Shu, L. (2025). Strain-engineered flexoelectricity and oxygen vacancy dynamics for modulating interfacial charge transport in BaTiO3. Journal of Materiomics, 101145.

[26] Lambán Jiménez, L. J., Jódar, J., Martos Rosillo, S., González Ramón, A., & Heredia Díaz, J. G. (2022). Assesment of climate change impact in mountain areas of special environmental value.

[27] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[28] Jin, F., Fan, S., Gu, M., Lv, Q., Ge, M., Zhang, Z., ... & Wu, W. (2025). Topotactic Reduction-Driven Crystal Field Excitations in Brownmillerite Manganite Thin Films. Advanced Functional Materials, 35(28), 2501893.

[29] Urrutia, E., Ayesta, I., Azkune, M., & Zubia, J. (2025). Recent advances in sensor applications of microstructured optical fibers: A review. IEEE Sensors Journal.

[30] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[31] Leong, C. Y., Cui, J., Cheng, X., Htein, L., & Tam, H. Y. (2026). Tactile-Sensitive Artificial Skin for Multiaxial Force Detection and Texture Recognition. Journal of Lightwave Technology, 44(2), 767-775.

[32] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[33] García Gil, A., Baquedano Estévez, C., & Garrido Schneider, E. (2022). The SAGE4CAN project: The use of shallow geothermal energy from the Canary Islands.

[34] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[35] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[36] Wang, S., Fu, X., Lv, J., Zhu, C., Zhang, G., & Meng, Y. (2025). Tailoring tribological properties in Titanium alloy via pulse Laser-induced textures: A review. Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology, 239(12), 1581-1602.

[37] Mejías Moreno, M. (2022). Research Group on Applied Hydrogeology and Shallow Geothermal Energy (GIHAGS): general information and research lines.

[38] Efe, I., Vogel, A., Huxter, W. S., Gradauskaite, E., Gaponenko, I., Paruch, P., ... & Trassin, M. (2025). Nanoscale electrostatic control in ferroelectric thin films through lattice chemistry. Nature Communications, 16(1), 6131.

[39] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[40] Carpenter, D. E., Nasiri, Z., Palagiri, N. R., Strickland, K. L., Harris, S. B., Geohegan, D. B., & Camata, R. P. (2026). Pulsed laser synthesis of mesoporous metal chalcogenide thin films. Applied Surface Science, 165842.

[41] Nedoma, J., Krizan, D., Stipal, J., Pereira, L., Bekmurzayeva, A., Tosi, D., ... & Marques, C. (2026). Decade of advancements in light–matter interaction-based optical fiber biosensing: innovations, challenges, and future directions. Advanced Photonics, 8(1), 014004-014004.

[42] Durán Valsero, J. J., Aguilera Alonso, H., Álvarez Alonso, R., Castro Dorado, A., Losa Román, A. D. L., Domínguez Sánchez, J. A., ... & Ruiz Hernández, J. M. (2022). Responding to the challenges of Water and Global Warming: Environmental Hydrogeology and Global Change Research Group (HYGLO-Lab).

[43] Tiwari, M. M., Noormohammed, S., Sarkar, D. K., & Chen, X. G. (2025). Surface treatment of PVC by corona discharged atmospheric air plasma for adhesive joining with AA 6061 aluminium alloy. Journal of Adhesion Science and Technology, 39(5), 734-749.

[44] Ikram, Z. (2025, May). Depth-Guided Neural Network for Robust Face Anti-Spoofing. In 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE.

[45] Pardo-Igúzquiza, E., Durán Valsero, J. J., Luque Espinar, J. A., Morales García, R., Heredia Díaz, J. G., Moreno Merino, L., ... & Balard, D. (2022). Hydrogeology and geomorphology of high relief karst systems: Sierra de las Nieves (Málaga) and Sierra de Tendeñera (Huesca).

[46] Pedrera Parias, A., Babault, J., Díaz Alvarado, J., García Senz, J. M., Hidas, K., López Mir, B., ... & Ruiz Constán, A. (2022). Tectonic Processes and GeoResources Lab.

[47] Ibrayev, S., Omarov, B., Amanov, B., & Momynkulov, Z. (2024). Development of a deep learning-enhanced lower-limb exoskeleton using electromyography data for post-neurovascular rehabilitation. Engineered Science, 31, 1269.

[48] Zukri, M. N. M., Al Farisi, M. S., Hasegawa, Y., & Shikida, M. (2025). Single-step laser patterning and thinning of biocompatible MEMS flow sensor. Sensors and Actuators A: Physical, 117377.

[49] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[50] Schaller, M., Hloch, S., Nag, A., Klichova, D., Pude, F., Zelenak, M., ... & Rosenhahn, B. (2025). Audio Monitoring of Bone Cement Disintegration in Pulsating Fluid Jet Surgery under Laboratory Conditions. Results in Engineering, 108762.

[51] Rajebahadur, S., Velluva Rayaroth, N., King, J. B., Alexandrovskaya, Y. M., Guagliumi, G., Glatz, A., ... & Sobol, E. N. (2025). Increasing arterial compliance by laser modification of fibro-calcific plaques. Frontiers in Cardiovascular Medicine, 12, 1652529.

[52] García Aróstegui, J. L. (2022). Innovative tools for the integrated management of groundwater in a context of increasing scarcity of water resources: Interreg Sudoe AQUIFER Project.

[53] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[54] Robador Moreno, A., Díez Fernández, R., Díez Montes, A., Gonzalo Guerra, B., Heredia, N., Mancebo Mancebo, M. J., ... & Rubio Pascual, F. J. (2022). Geological mapping as a tool for scientific research: Group of Geological, Geomorphological and 4D Geology Cartography of Mountain Ranges and Basins.

[55] Sultan Mukhamedaly, Kymbat Kabekeyeva, Gulnar Mussabekova, Aliya Kuralbayeva, Bagdat Toibekova, Gulzhan Makashkulova, Batyrkhan Omarov, "A Pedagogical Framework for Ethical Skill Development in Higher Education within Smart Learning Environments", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 1-20, 2026. DOI:10.5815/ijmecs.2026.02.01

[56] Ikram, Z., & Omarov, B. (2024). Fusion of vision transformers and convolutional networks for advanced face anti-spoofing. In DTESI.

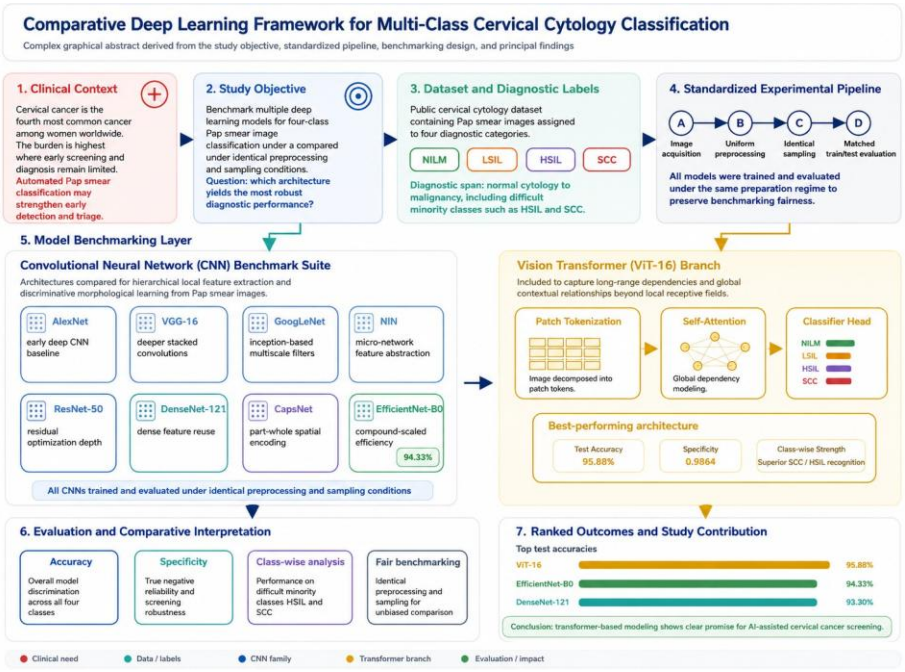
[57] Jiménez Sánchez, J., Meléndez Asensio, M., Peinado Parra, T., & Fernández Jurado, M. (2022). Eventual effect on groundwater in the hydrographic demarcation of La Palma due to the eruption of the Cumbre Vieja volcano.

Cervical Cytology Classification Using Multiple CNN Architectures with Transformer-Based Feature Enhancement

Mehreen Sirshar[1*], Omama Shakeel[2], Nayab Asim[3], Fakeeha Jafari[4],
Hani Almoamari[5], Adnan Nadeem[6], Mohammad Zubair Khan[7*], Ibrahim Aljubayri[8]
Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan [1, 2, 3]
Department of Software Engineering, National University of Modern Languages, Pakistan [4]
Faculty of Computer and Information System, Islamic University of Madinah, Madinah 42351, Saudi Arabia [5, 6, 7]
Department of Computer Science and Information,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia [8]

***Abstract*—Cervical cancer is the fourth most common cancer among women worldwide and remains a major public health challenge, particularly in regions with limited access to early screening and diagnosis. Accurate classification of cervical cytology images is critical for early detection of cervical cancer, which remains a major health burden in low- and middle-income countries. This study presents a comprehensive evaluation of multiple deep learning architectures for the automated classification of Pap smear images into four diagnostic categories: Negative for Intraepithelial Lesion or Malignancy (NILM), Low-Grade Squamous Intraepithelial Lesion (LSIL), High-Grade Squamous Intraepithelial Lesion (HSIL), and Squamous Cell Carcinoma (SCC). We systematically compare eight Convolutional Neural Network (CNN) architectures: AlexNet, VGG-16, GoogLe-Net, Network-in-Network (NIN), ResNet-50, DenseNet-121, Capsule Networks, and EfficientNet-B0 on a publicly available cervical cytology dataset. To enhance feature representation and capture long-range dependencies, we additionally incorporate a Vision Transformer (ViT-16) model. All models are trained and evaluated under identical preprocessing and sampling conditions to ensure fair benchmarking. Experimental results demonstrate that ViT-16 achieves the highest test accuracy of 95.88% and an overall specificity of 0.9864, outperforming all CNN counterparts. EfficientNet-B0 and DenseNet-121 also showed strong performance, achieving 94.33% and 93.30% accuracy, respectively. Notably, ViT-16 provided superior classification outcomes for challenging minority classes such as SCC and HSIL. The findings highlight the growing potential of transformer-based models in cytopathology and underscore the importance of architectural design in developing robust diagnostic tools. This work contributes a comparative foundation for future research in AI-assisted cervical cancer screening systems.**

***Keywords*—Cervical cancer; CNN architectures; vision transformer**



Graphical abstract.

*Corresponding author.

I. Introduction

The cervix is the lower, narrow portion of the uterus that connects the uterine cavity to the vaginal canal. Functionally, it plays a vital role in female reproductive health by facilitating the passage of sperm, acting as a barrier to infections, and serving as the birth canal during labor. Despite its biological importance, the cervix is susceptible to malignant transformations that lead to cervical cancer.

Cervical cancer arises primarily due to persistent infection with high-risk human papillomavirus (HPV) types. According to Delgado 2022 [1], cervical cancer accounted for approximately 604,000 new cases and 342,000 deaths worldwide, making it the fourth most common cancer among women globally. The disease burden is disproportionately higher in low- and middle-income countries, where access to early screening and preventive healthcare remains limited. These statistics highlight the urgent need for reliable and scalable diagnostic systems to support early detection and intervention. The disease progresses through precancerous stages, including low-grade squamous intraepithelial lesions (LSIL) and high-grade squamous intraepithelial lesions (HSIL), before developing into invasive squamous cell carcinoma (SCC). Early detection and classification of these lesions are crucial for improving survival rates and reducing treatment complexity. Fig. 1 shows an overview of cervical cytology classes: (A) HSIL, (B) LSIL, (C) NILM, and (D) SCC Pap smear images representing different diagnostic categories.

Traditionally, cervical cytology diagnosis is performed using the Papanicolaou (Pap) smear test, which relies on expert microscopic examination of cervical epithelial cells. Although effective, manual screening is labor-intensive, subjective, and prone to inter-observer variability. In response to these challenges, artificial intelligence (AI) and deep learning approaches have gained traction for automated image-based diagnostics. Convolutional Neural Networks (CNNs), in particular, have demonstrated superior performance in feature extraction and classification tasks across medical imaging domains [2, 3].

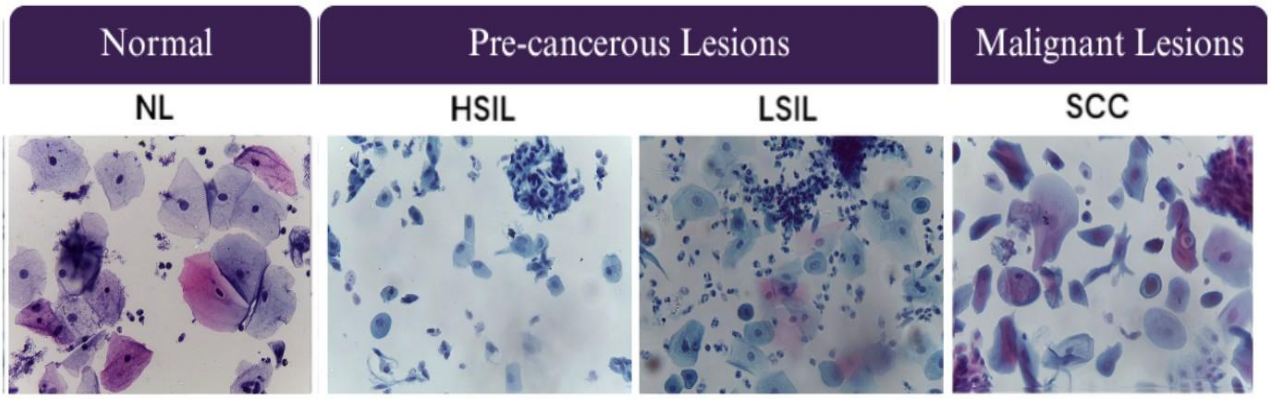


Fig. 1. Overview of cervical cytology classes: (A) HSIL, (B) LSIL, (C) NILM, and (D) SCC pap smear images representing different diagnostic categories.

Recent studies have introduced various CNN-based architectures for cervical cancer detection. For instance, Liu et al. [4] proposed a hybrid CNN and Transformer model achieving high accuracy in classifying Pap smear images. Similarly, Gangrade et al. [5] and Huang et al. [3] utilized ensemble and fine-tuned deep learning techniques to improve classification outcomes. These models outperform conventional machine learning algorithms by reducing manual feature engineering and enhancing generalizability [6].

Nevertheless, issues such as class imbalance, image variability, and staining artifacts continue to challenge model robustness and reliability [8, 9]. To address these, researchers have incorporated data augmentation [6] and attention-based models [10] to enhance feature learning. Despite such advancements, the evaluation of different CNN architectures on a balanced and diverse dataset remains a key research gap.

Cervical cytology classification via deep options has been an active area of research. Many studies rely primarily upon transferring knowledge from other domains (transfer learning) and focus on establishing high levels of performance within ideal conditions. Very few published studies systematically examine multiple convolutional neural network (CNN) architectures and/or transformers under standardized training conditions and/or real, imbalanced, cervical cytology datasets. Therefore, some controlled comparative experiments will be performed to determine the impacts of architectural designs on robustness of the classification under the constraints of realistic clinical data.

The remainder of this study is organized as follows: Section II presents the related work with a focus on recent advances in cervical cancer classification using deep learning. Section III details the Materials and Methods, model architectures, and Dataset. Section IV discusses the experimental setup, Methodology, Training, and Implementation Details, and Section V provides performance comparisons among the evaluated models. Finally, Sections VI and VII present the Discussion and Conclusion that outline potential directions for future research.

II. Related Work

Cervical cancer remains a leading cause of mortality among women worldwide, emphasizing the urgent need for early and accurate diagnostic tools. Deep learning techniques, particularly convolutional neural networks (CNNs) and transformer-based models, have demonstrated strong potential in automating

cervical cytology image classification. This section critically reviews recent studies, analyzing their methodological innovations, performance outcomes, and relevance to real-world medical applications.

Nirmala et al. [2] introduced an adaptive system combining CNNs with adaptive vision transformers, achieving 97.43% accuracy through refined segmentation and classification strategies. Similarly, Huang et al. [3] employed AF-SENet with fine-tuned CNNs to enhance the detection of LSIL, pushing the accuracy to 95.33%. Liu et al. [4] proposed the CVM-Cervix framework, which synergizes CNNs, visual transformers, and MLPs, attaining 91.72% accuracy and exemplifying the advantage of hybrid architectures.

Several researchers explored ensemble and multi-model strategies. Gangrade et al. [5] combined CNN, AlexNet, and SqueezeNet, achieving 94% accuracy in five-class classification, while Tripathi et al. [6] applied majority voting across CNN models to improve classification reliability. Benyes et al. [6] used AE-CNN on SurePath and ThinPrep datasets, reporting 96.54% accuracy and demonstrating effective domain adaptation.

Signal processing integration was explored by Palanisamy et al. [7], who combined dual-tree complex wavelet transform (DTCWT) with CNNs, achieving 99% accuracy across four Pap smear categories. Complementarily, Sellamuthu et al. [7] developed a CNN pipeline with augmentation and pooling, achieving a Pap smear detection index of about 99%.

Transfer learning techniques also showed promise. Tan et al. [8] tested 13 pretrained CNNs and identified DenseNet-201 as the most effective on the Herlev dataset. Novitasari et al. [9] applied ResNet-50 to classify colposcopy images into five stages with 100% accuracy, while Wulandari et al. [9] highlighted the stability of ResNets in mitigating vanishing gradient problems.

Whole-slide image (WSI) analysis was addressed by Cheng et al. [10], who proposed a progressive lesion recognition system using dual-resolution WSIs, achieving 95.1% sensitivity and 93.5% specificity. Kaur et al. [11] compared 16 CNN models and reported 95% accuracy with ResNet-50 and 99.95% with VGG16 for binary classification, emphasizing the strength of pre-trained models.

Alsubai et al. [12] integrated PCA and GWO for feature selection in a CNN-based classifier on SIPaKMeD data, achieving 91.13% accuracy and strong precision. De La Cruz Paucar et al. [13] improved generalization by employing dropout in a two-stage CNN, where ResNet-50 outperformed competing models. Finally, Zhao et al. [14] tackled data imbalance by generating cervical images using taming transformers, thereby enhancing classification accuracy and dataset diversity. Recent advances in medical image classification and deep learning architectures, including transformer-based models and conventional CNN frameworks, have demonstrated strong potential for improving automated diagnostic systems in medical imaging applications [15–22]. Recent studies have also explored optimization-driven and hybrid deep learning approaches for medical image analysis and disease classification tasks [31–33].

These studies collectively underline the transformative impact of deep learning on cervical cancer screening, while also revealing ongoing challenges related to class imbalance, model generalizability, and clinical-scale deployment. Table I shows a comparative summary of related studies on cervical image analysis.

III. Materials and Methods

This section outlines the dataset, preprocessing techniques, model architectures, and training strategies used in this study. The proposed methodology is designed to classify cervical cytology images into four diagnostic categories using multiple CNN architectures and a transformer-based feature enhancement stage. The process includes image preprocessing, CNN-based feature extraction, integration of a transformer model for contextual learning, and final classification.

A. Dataset

This study employs the publicly available Mendeley cervical cancer image dataset [23] (DOI:10.17632/zddtpgzv63.4), which includes high-resolution cytology slide images stained using the Papanicolaou (Pap) method. The dataset is composed of 963 images representing four clinically annotated classes: Negative for Intraepithelial Lesion or Malignancy (NILM), Low-grade Squamous Intraepithelial Lesion (LSIL), High-grade Squamous Intraepithelial Lesion (HSIL), and Squamous Cell Carcinoma (SCC). Images were captured at 400× magnification (using a Leica ICC50 HD attached to a microscope) and sourced from 460 patients in real-world clinical settings. Each image in the dataset has been labeled by expert cytologists. The dataset presents real-world variability in terms of staining artifacts, cell overlap, and illumination differences, making it a suitable benchmark for evaluating the robustness of deep learning models. However, it suffers from class imbalance, with the NILM class being overrepresented and SCC and LSIL being underrepresented. This imbalance poses a challenge for classifiers and is addressed through class-weighting and sampling strategies in training.

B. Computing Infrastructure

The Experiments are conducted in Google Colab using a standard Linux-based runtime environment. The operating system is a 64-bit Linux distribution provided by the Colab virtual machine. Model development and training were implemented in Python using the PyTorch deep learning framework with supporting scientific libraries (NumPy, Matplotlib, and scikit-learn). Experiments are executed under Colab's managed hardware configuration, and where GPU acceleration was available, training leveraged an NVIDIA CUDA-enabled GPU provided by the platform. In cases where GPU is not enabled, training is performed on the Colab CPU backend. Data are accessed from cloud storage (Google Drive) and loaded using a standard folder-based dataset structure compatible with ImageFolder. All pre-processing (resizing, normalization, and tensor conversion) and evaluation (confusion matrix, ROC/AUC, and classification report) are performed within the same environment to ensure consistency. The same runtime setup is used across all model variants to maintain comparable experimental conditions.

TABLE I. COMPARATIVE SUMMARY OF RELATED STUDIES ON CERVICAL IMAGE ANALYSIS

Study	Year	Techniques	Application	Strengths	Weaknesses
Nirmala et al. [2]	2025	CNN + Adaptive ViT	Pap smear classification	High accuracy (97.43%); robust segmentation and classification	Sensitive to data imbalance and segmentation quality
Huang et al. [3]	2020	AF-SENet + fine-tuned CNN	LSIL recogni-tion	Improved LSIL detection; effective deep feature fusion	Requires careful tuning; limited generalizability
Liu et al. [4]	2022	CNN + ViT + MLP (CVM-Cervix)	Pap smear classification	Hybrid synergy; solid accuracy (91.72%)	Complex model; dataset-dependent benefits
Gangrade et al. [5]	2025	Ensemble (CNN, AlexNet, SqueezeNet)	Five-class cervical classification	Ensemble outperforms single models (94%)	Higher inference cost; requires model aggregation
Tripathi et al. [6]	2022	Multiple CNNs + majority voting	Multiclass Pap smear	Voting improved reliability	Still sensitive to class imbalance
Benyes et al. [6]	2022	AE-CNN + domain adaptation	Pap smear (SurePath/Thin-Prep)	Strong accuracy (96.54%); cross-prep robustness	Dataset-specific adaptation; limited minority class detail
Palanisamy et al. [7]	2022	DTCWT + CNN	Pap smear (4 categories)	Very high accuracy (99%); DL + signal synergy	Overfitting risk; preprocessing complexity
Sellamuthu et al. [7]	2022	CNN + augmentation + pooling	Pap smear detection index	Robust with augmentation; simple pipeline	Limited innovation; weak for rare classes
Tan et al. [8]	2024	TL with 13 pretrained CNNs (DenseNet-201 best)	Pap smear (Her-lev)	Effective under data scarcity	Dependent on pre-train domain
Novitasari et al. [9]	2022	ResNet-50	Colposcopy, 5 stages	Reported 100% accuracy; strong residual learning	Possible overfitting; domain-specific
Wulandari et al. [9]	2022	Residual networks (multi-stage)	Colposcopy	Mitigates vanishing gradients	Requires high compute; dataset dependency
Cheng et al. [10]	2021	Progressive lesion recognition (WSIs)	Whole-slide imaging	Clinically relevant; high sensitivi-ty/specificity	Compute-heavy; pipeline complexity
Kaur et al. [11]	2025	16 pretrained CNNs (ResNet-50, VGG16)	Pap smear 7	Strong baselines (VGG16: 99.95% binary)	Binary results may not generalize to multiclass
Alsubai et al. [12]	2023	PCA + GWO + CNN	SIPaKMeD dataset	Good accuracy (91.13%); feature reduction helps	Risk of losing rare-class fea-tures
De La Cruz Paucar et al. [13]	2024	Two-stage CNN + dropout (ResNet-50 best)	Pap smear classification	Dropout improved generalization	Limited external validation
Zhao et al. [14]	2022	Taming transformers (data generation)	Cervical image augmentation	Tackles imbalance; boosts accuracy/diversity	Synthetic–real domain gap; validation needed

IV. EXPERIMENTAL SETUP

This section outlines the practical implementation of the proposed methodology, including model training protocols, evaluation metrics, and analysis tools. Each component of the setup, from data preprocessing to final evaluation, is described below.

A. Training and Implementation Details

All of the models have been created with the same dataset structure and preprocessing pipeline. The data preprocessing for images involved resizing all of the images to one common size matching each of the architectures, converting all images to tensors, and normalizing all images with the mean and standard deviation of ImageNet. To keep the experiments reproducible, we pre-split our dataset into three parts (70% training, 15% validation, and 15% testing) using a stratified split. The stratified split preserved the class representation in the training, validation, and testing datasets for all four classes of diagnosis (NILM, LSIL, HSIL, SCC). Random number generator seed 42 was used to split the dataset and create and initialize the models to ensure consistent conditions in these experiments. To handle the great majority of the imbalanced dataset, class weightings were determined based on the number of occurrences of each class in the training set, and we utilized a WeightedRandomSampler during training to create balanced mini-batches of samples. Balanced mini-batches allowed the SCC class to receive better representation in terms of class size and ultimately provided better representation of the SCC class' actual proportion of the study sample.

The Adam optimizer is used for each of the models at a learning rate of 0.0001 and a batch size of 32. The weighted cross-entropy loss function is utilized to incorporate class weights into the optimization process. Each model is trained for 10 epochs, with training and validation loss and accuracy being used to monitor model performance.

The preliminary experiments showed that model validation performance plateaus around 8–10 epochs, indicating early convergence given the moderate size of the dataset. Training models beyond these epochs did not show large increases in model performance and created a higher risk of overfitting; therefore, all models were trained for 10 epochs to maintain consistency and to not over-optimize the models. After training, the best-performing model checkpoint (based on validation accuracy) is evaluated on the independent test set. All experiments are conducted under the same computational environment to ensure comparability across architectures.

B. Evaluation Metrics

The following equations define the metrics used to evaluate model performance:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

Area Under the ROC Curve (AUC): Measures the model's ability to distinguish between classes across all thresholds.

Specificity Analysis: Class-wise specificity is calculated to assess each model's ability to correctly identify negative samples. Overall specificity is derived by averaging across classes.

In addition to class-wise metrics, the final test accuracy and average specificity are computed to summarize overall model effectiveness. These metrics provide both class-wise and overall assessments of the model's classification ability, particularly highlighting performance on underrepresented classes such as *SCC*.

C. Methodology

The proposed methodology is designed to classify cervical cytology images into four diagnostic categories using multiple CNN architectures and a transformer-based feature enhancement stage. The process includes image preprocessing, CNN-based feature extraction, integration of a transformer model for contextual learning, and final classification.

1) Deep learning models employed: In this study, we systematically evaluate eight distinct CNN architectures—AlexNet, VGGNet, GoogLeNet, Network-in-Network (NIN), ResNet, DenseNet, Capsule Networks, and EfficientNet—for cervical cytology classification. The objective is to compare their classification performance and assess their capability in addressing limitations such as computational inefficiency, poor spatial representation, and vanishing gradients. By using a publicly available cytology image dataset, we aim to provide a comprehensive benchmark and insights into architecture selection for real-world AI-assisted cervical cancer diagnostics.

AlexNet (2012) serves as a foundational CNN model, comprising five convolutional layers and three fully connected layers, totaling approximately 61 million parameters. Though relatively shallow, it provides a strong baseline for evaluating improvements in newer models [24]. Advantages: Simple and efficient for moderate datasets. Limitations: Limited depth reduces capability for complex feature learning.

VGGNet (2014) introduces a deeper network with about 138 million parameters using uniform 3×3 convolution filters. This architecture improves hierarchical feature learning but is computationally expensive [30]. Advantages: Effective in learning fine-grained features. Limitations: High training time and memory demand. Motivation: To analyze the benefit of increased depth over AlexNet's simpler design.

GoogLeNet (Inception) (2014) features inception modules for multi-scale feature extraction and contains roughly 6.8 million parameters. Its parallel convolution paths make it both deep and computationally efficient [26]. Advantages: Learns hierarchical and multi-scale features. Limitations: Complex to tune and modify. Motivation: Introduced to evaluate multi-path processing and spatial efficiency.

Network in Network (NIN) (2014) replaces dense layers with micro neural networks, namely MLPs, for increased abstraction [29]. It utilizes about 29 million parameters. Advantages: Improved local abstraction through MLPs. Limitations: Lacks the expressiveness of deeper residual connections. Motivation: Explores local abstraction techniques in contrast to global dense features.

ResNet (2015) addresses the vanishing gradient problem through skip connections. ResNet-50, with 25.6 million parameters, enables training of very deep networks [25]. Advantages: Residual learning enables depth without degradation. Limitations: Computationally demanding despite a moderate parameter count. Motivation: Evaluates the effectiveness of residual connections in classification performance.

DenseNet (2017) connects each layer to all subsequent ones to encourage feature reuse and gradient flow [27]. It contains around 8 million parameters. Advantages: Feature reuse, reduced parameters, and strong gradient propagation. Limitations: Memory usage can increase due to feature map concatenation. Motivation: Introduced to compare dense connectivity with residual learning.

EfficientNet (2019) balances depth, width, and resolution using compound scaling. With only 5.3 million parameters, EfficientNet-B0 offers high accuracy at low computational cost [28].

Advantages: Lightweight and efficient architecture. Limitations: May underperform on extremely complex or high-resolution inputs. Motivation: Explores whether efficient models can achieve performance comparable to deeper networks.

Vision Transformer (ViT) (2021) enhances feature representation by applying transformer architecture to image patches [16]. Each image is split into fixed-size patches, linearly embedded, and processed by a stack of transformer encoders using multi-head self-attention. We fine-tuned a pretrained ViT-B/16 model on our dataset. Advantages: Captures long-range dependencies and global context. Limitations: Requires large datasets or pretraining for optimal performance. Motivation: Introduced as a transformer-based alternative to CNNs for global feature learning in cytology images. Fig. 2 shows an overview of the CNN architectures used in this study. Fig. 3 shows the architecture of the Vision Transformer (ViT). The image is divided into fixed-size patches, which are embedded and processed through multiple transformer encoder blocks with multi-head self-attention before final classification.

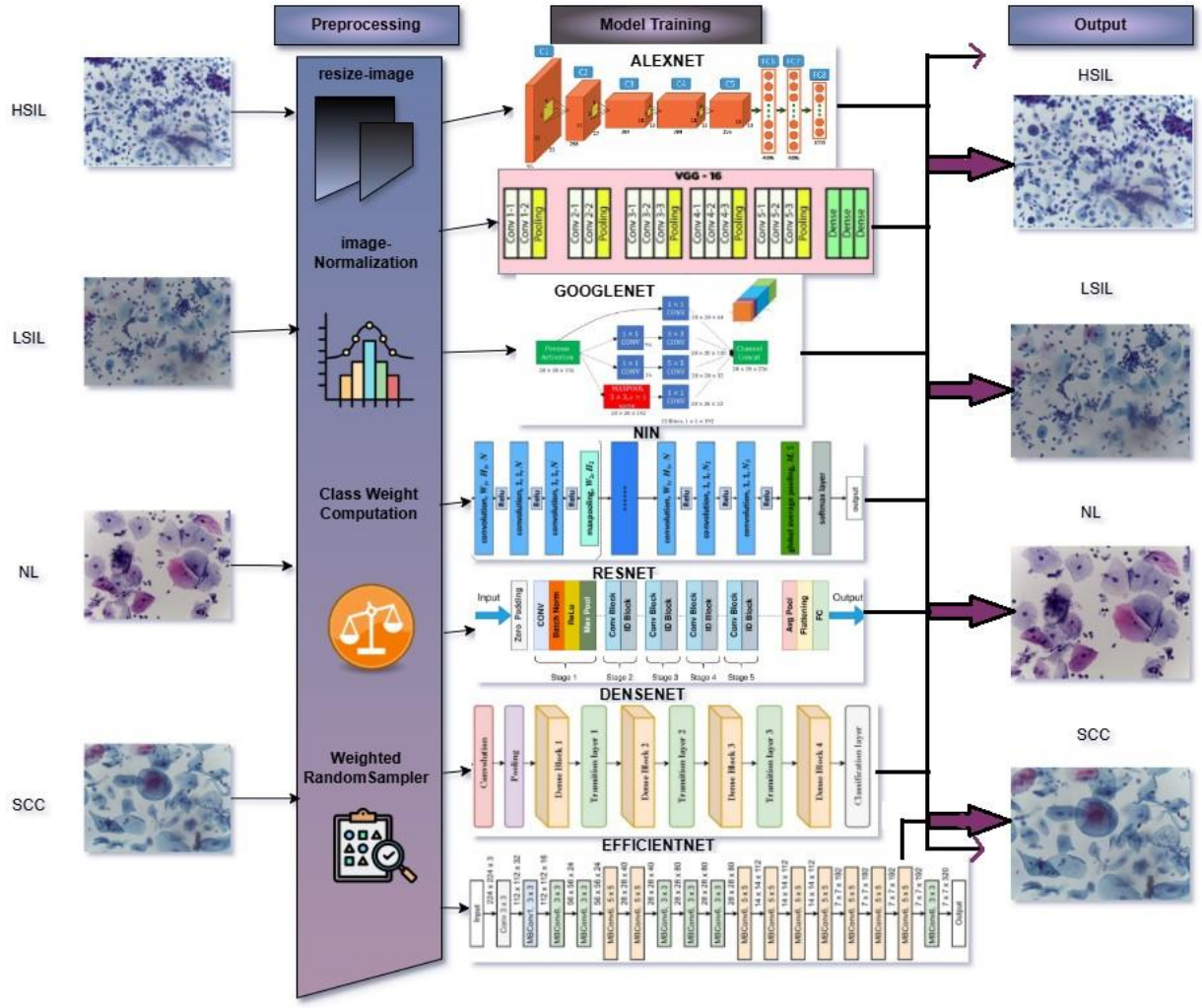


Fig. 2. Overview of CNN architectures used in this study: (A) AlexNet [24], (B) VGGNet [30], (C) GoogLeNet [26], (D) Network-in-Network (NIN) [29], (E) ResNet [25], and (F) DenseNet [27]. These architectures were evaluated to classify cervical cytology images into four diagnostic categories.

V. Evaluation Method

In this section, we present and analyze the classification performance of eight CNN architectures trained without fine-tuning over 10 epochs using a batch size of 32. Each model is evaluated on the same cervical cytology dataset using multiple performance metrics, including accuracy, precision, recall, F1-score, and specificity across four diagnostic categories: HSIL, LSIL, NMIL, and SCC. The dataset is split into training, validation, and test sets, with test results forming the basis of this evaluation to ensure consistent comparison across models.

A. Model-Wise Performance Analysis

The comparative performance analysis of the models shows varying levels of effectiveness. AlexNet achieved a final test accuracy of 66.49% with a specificity of 0.9018, performing relatively well for HSIL recall (0.94) but showing low effectiveness for SCC (recall: 0.07). ResNet-50 improved the results with an accuracy of 81.96% and specificity of 0.9499, offering strong recall for SCC (0.93) along with balanced performance across the other classes. GoogLeNet reported 81.44% accuracy, providing high precision and recall for LSIL and NMIL but moderate performance for SCC. DenseNet-121 reached 93.30% accuracy and 0.9748 specificity, performing consistently across all categories, including the minority class SCC. EfficientNet-B0 achieved 94.33% accuracy with a specificity of 0.9832, maintaining stable precision and recall across diagnostic groups. Network-in-Network (NIN) showed relatively lower performance, with 69.59% accuracy and 0.8237 specificity, and had difficulties with LSIL and SCC. VGG-16 performed with 90.72% accuracy and 0.9725 specificity, showing reliable results for HSIL, LSIL, and NMIL but lower recall for SCC. Finally, Vision Transformer (ViT16) obtained the highest accuracy of 95.88% and specificity of 0.9864, with improved recall for SCC and consistent overall classification. Fig. 4 and 5 show confusion matrices and ROC curves for CNN models. Table II present Comparative performance of pretrained deep learning models on the cervical cytology dataset. All models were evaluated under identical training and preprocessing conditions. Table III presents per-class Precision (P), Recall (R), and F1-score for all evaluated pretrained models on the cervical cytology dataset. Fig. 6 shows ROC curves of the ViT-16 model for four-class cervical cytology.

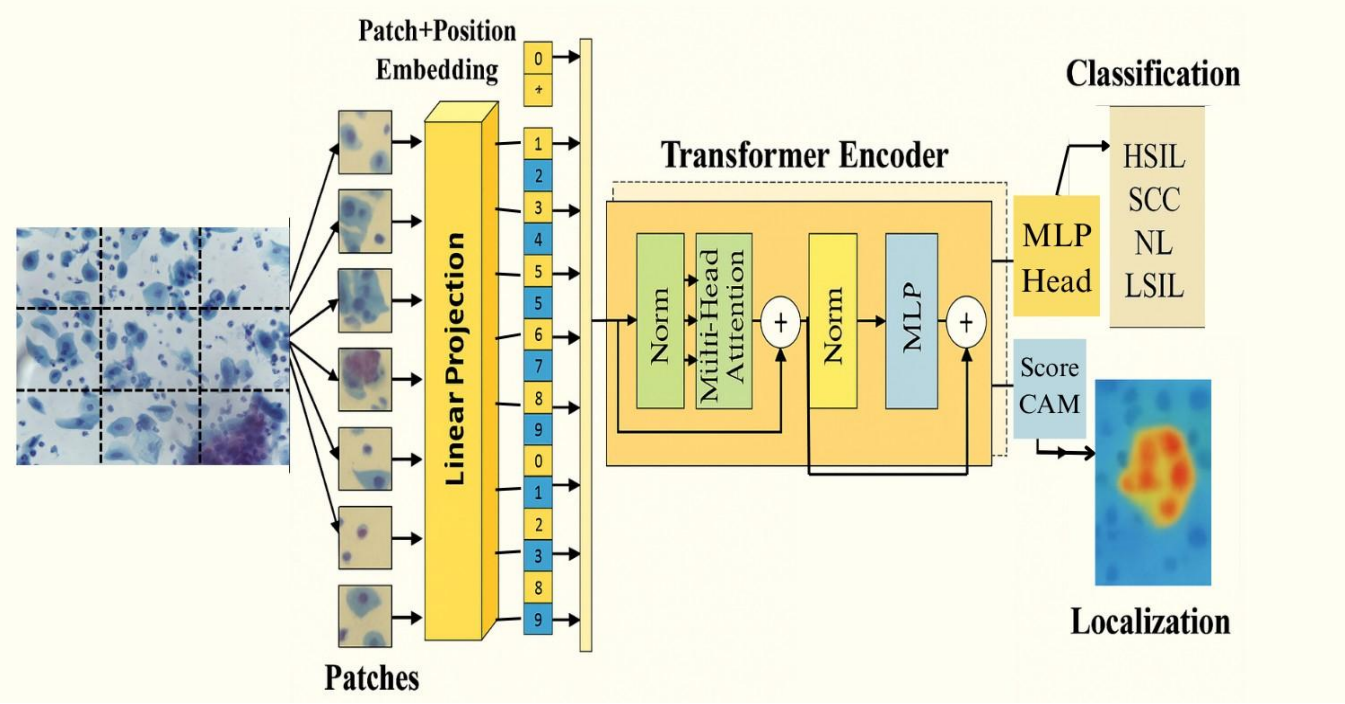


Fig. 3. Architecture of the Vision Transformer (ViT). The image is divided into fixed-size patches, which are embedded and processed through multiple transformer encoder blocks with multi-head self-attention before final classification.

TABLE II. COMPARATIVE PERFORMANCE OF PRETRAINED DEEP LEARNING MODELS ON THE CERVICAL CYTOLOGY DATASET. ALL MODELS WERE EVALUATED UNDER IDENTICAL TRAINING AND PREPROCESSING CONDITIONS

Model	Accuracy (%)	Macro F1	SCC Recall	Overall Specificity	Best For
AlexNet	66.49	0.57	0.07	0.9018	HSIL Recall
ResNet-50	81.96	0.70	0.93	0.9499	SCC Detection
GoogLeNet	81.44	0.76	0.53	0.9463	LSIL Classification
DenseNet-121	93.30	0.86	0.73	0.9748	Balanced Performance
EfficientNet-B0	**94.33**	0.84	0.40	0.9832	High Overall Accuracy
Network-in-Network (NiN)	69.59	0.35	0.00	0.8237	HSIL/NILM Detection
VGG-16	90.72	0.75	0.13	0.9725	LSIL/NILM Classification
ViT-B/16	**95.88**	**0.89**	**0.87**	**0.9864**	Overall Best Performance

TABLE III. PER-CLASS PRECISION (P), RECALL (R), AND F1-SCORE FOR ALL EVALUATED PRETRAINED MODELS ON THE CERVICAL CYTOLOGY DATASET

Model		HSIL			LSIL			NILM			SCC	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AlexNet	0.52	0.94	0.67	0.68	1.00	0.81	0.99	0.60	0.75	0.04	0.07	0.05
ResNet-50	0.73	0.33	0.46	0.72	1.00	0.84	1.00	0.90	0.95	0.39	0.93	0.55
GoogLeNet	0.58	0.91	0.71	1.00	1.00	1.00	1.00	0.79	0.88	0.36	0.53	0.43
DenseNet121	0.87	0.82	0.84	1.00	0.87	0.93	0.98	1.00	0.99	0.65	0.73	0.69
EfficientNetB0	- 0.78	0.94	0.85	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.40	0.52
Network-in-Network (NIN)	0.54	0.58	0.56	0.00	0.00	0.00	0.73	0.94	0.82	0.00	0.00	0.00
VGG-16	0.69	1.00	0.81	0.96	1.00	0.98	1.00	0.96	0.98	0.50	0.13	0.21
ViT-B/16	**0.93**	**0.85**	**0.89**	**1.00**	0.96	**0.98**	0.99	**1.00**	**1.00**	**0.72**	**0.87**	**0.79**

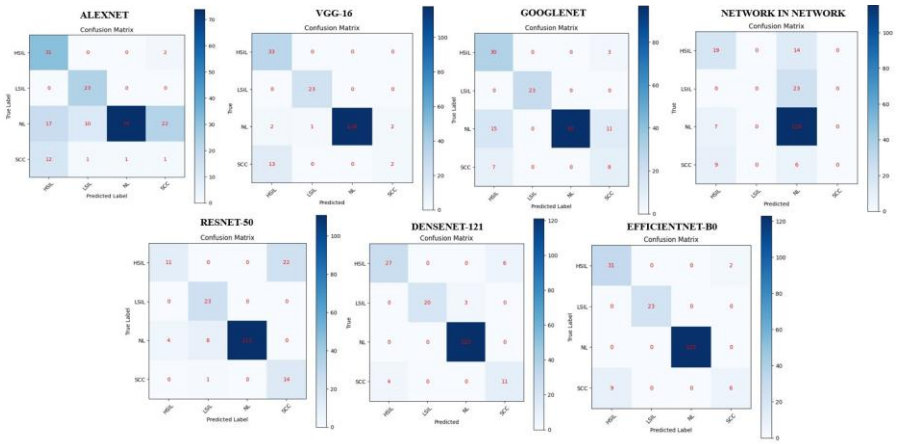


Fig. 4. Confusion matrices for CNN models including DenseNet-121, NIN, AlexNet, GoogLeNet, ResNet-50, VGG-16, and EfficientNet-B0 on the cervical cytology dataset.

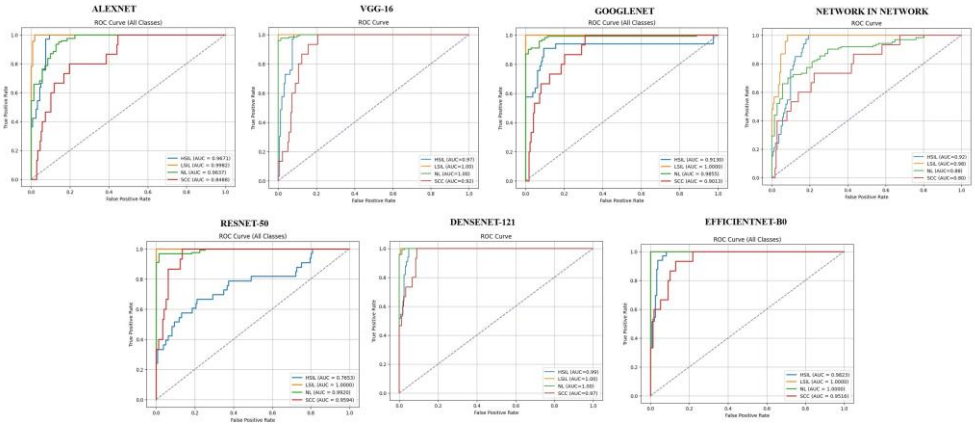


Fig. 5. ROC curves for CNN models including AlexNet, VGG-16, GoogLeNet, NIN, ResNet-50, DenseNet-121, and EfficientNet-B0 on the cervical cancer dataset.

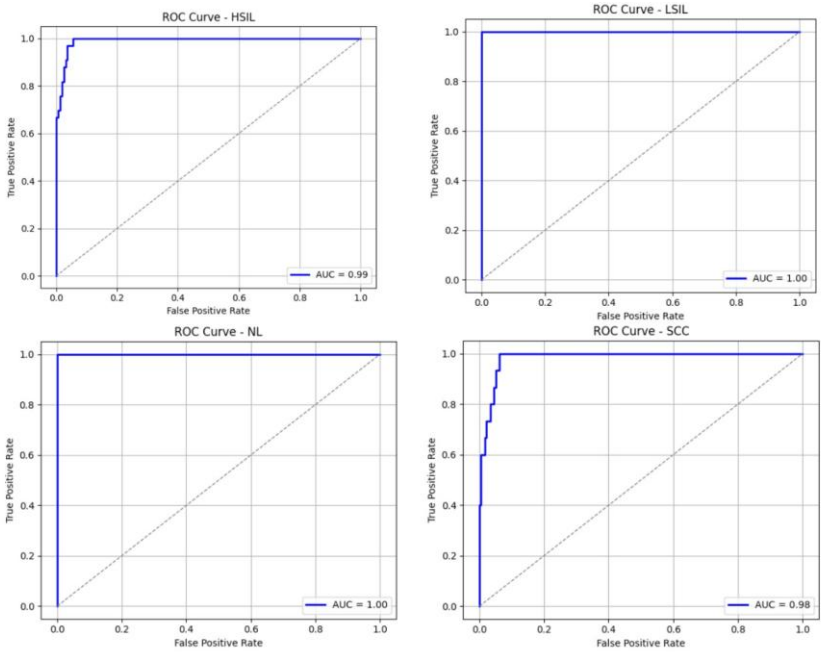


Fig. 6. ROC curves of the ViT-16 model for four-class cervical cytology classification: HSIL, LSIL, NILM, and SCC. The model achieved high AUC values of 0.99, 1.00, 1.00, and 0.98, respectively.

B. Best Model Overview

Among all models evaluated, EfficientNet-B0 and Vision Transformer (ViT-16) demonstrated the highest classification performance. EfficientNet-B0 achieved a final test accuracy of 94.33% with an overall specificity of 0.9832. ViT-16 further surpassed these results with a test accuracy of 95.88% and specificity of 0.9864. Table IV presents ViT-16's class-wise metrics.

TABLE IV. PERFORMANCE METRICS BY CLASS FOR VIT-16

Class	Precision	Recall	F1-score	Specificity
HSIL	0.93	0.85	0.89	0.9876
LSIL	1.00	0.96	0.98	1.0000
NMIL	0.99	1.00	1.00	0.9859
SCC	0.72	0.87	0.79	0.9721

VI. DISCUSSION

The in-depth study here provides information about the experimental results, connections between models, dataset characteristics, and challenges faced in evaluating results. This section goes into deeper detail about the important findings from the comparative study of CNN and transformer-based models on the classification of cervix cytology. The cervix cytology dataset used in this study consists of four cervix cytology classifications: HSIL high-grade squamous intraepithelial lesion, LSIL low-grade squamous intraepithelial lesion NILM (Negative for intraepithelial lesion or malignancy), and SCC (Squamous cell carcinoma). Each classification is different in how many total images are available for training, as well as how complex the images are visually. The cervix cytology dataset also presented several challenges, such as imbalanced class distributions, variability between images, inconsistent staining, and subtle visual differences between the HSIL and LSIL classes. The SCC class had the lowest number of training sample images when compared to the other three cervix cytology classifications, and therefore, many of the models had difficulty generalizing well to the SCC. With only 963 total images in this study, transfer learning with pre-trained models was used to reduce the chance of overfitting and to increase generalization ability for features. Weighted class sampling was also used during the training phase along with applying class-balanced loss to address issues related to imbalanced class distributions. The experiments also show that EfficientNet-B0 and ViT-B16 performed well enough in terms of both recall and specificity on the SCC class while demonstrating a reasonable amount of resilience against the data scarcity issue. The preprocessing steps like cropping, normalizing, and augmenting datasets are critical to ensure that all models are trained on the same, equalized, and enhanced dataset. Additionally, using a WeightedRandomSampler helped balance the number of samples from minority classes by creating balanced mini-batches for training and increasing SCC detection accuracy. These results provide evidence that using class balancing approaches improves learning stability and sensitivity for minority-class samples. The hyperparameter settings used to fine-tune the models are identical to one another and included, using pretrained ImageNet weights, the same preprocessing methods for all models, a batch size of 32, Adam optimizer (learning rate 0.0001), weighted cross-entropy loss, and a maximum number of 10 epochs of training. Using standardized hyperparameter settings provided a consistent basis for an unbiased comparison of all models. Initial experiments showed that validation accuracy converged around epochs 8-10 for the majority of models, and further training yielded little or no improvement in the model's classification accuracy, with an increased probability of overfitting.

All models differ in their respective performance based on the same conditions of training, with substantially different performances. The least productive performers in regard to both recall and specificity (regardless of which dataset) are the shallow networks, including both AlexNet and Network-in-Network (NIN) networks, with especially poor performance seen with respect to the SCC class. One potential reason for their limitations includes fewer feature extractions and less ability to capture a complex spatial representation. Conversely, as the models become deeper and more connected (DenseNet-121 and EfficientNet-B0), both the amount of dense feature reuse and compound scaling contributed to increased performance through 1) dramatically improving feature representation and 2) increasing classification consistency among all diagnostic categories. However, as the models become deeper, they also require greater computational complexity and money to train, which could limit their ability to deploy in real-time clinical environments that are limited in available resources. Among the models analysed, the transformer-based ViT-B16 had the highest overall classification score out of all models on all datasets. The self-attention mechanism of this model provided the capability to effectively model long-range contextual dependencies and global structural information, allowing the model to be more effective at distinguishing visually similar lesion classes. Additionally, the strong performance displayed by ViT-B16 when trained with a very limited amount of training data is likely due to factors such as transfer learning, standardisation of preprocessing, balanced sampling, and the representational power of the attention mechanism of transformers. These findings indicate the emerging promise of transformer-based methods in the medical imaging field. Multiple metrics were assessed, including accuracy, precision, recall, F1 score, specificity, and ROC-AUC analysis, and while overall accuracy provides a high-level representation of how well a model performs, it can sometimes lead to inaccurate conclusions about models trained on imbalanced medical datasets. For this reason, class-level metrics were used to measure the ability of the models to detect the minority class. The importance of specificity was emphasized as it measures the model's ability to predict negative results while attempting to detect positive results in clinical screening situations where false positives present an additional complication to the costly and sometimes dangerous process of diagnosing patients. EfficientNet-B0 and ViT-B16 were able to generate consistently good specificity estimates across all classes, including SCC, indicating excellent diagnostic utility. Both macro and weighted F1 scores supported that the models were generating consistent predictions across majority and minority classes. EfficientNet-B0 and ViT-B16 demonstrated superior performance on all of these metrics, which would indicate they would be able to generalise better than the other models considered and have balanced learning behaviour. The evaluation of ROC-AUC

analysis provided additional support for the robust discriminatory ability of the models across all eleven diagnostic classifications.

Architectural depth, dense connectivity, transfer learning and attention mechanisms are important in enhancing performance for the classification of cervical cytology specimens. Although traditional convolutional neural networks (CNN) architectures are still valid benchmark models, ViT (Vision Transformer) architectures such as ViT-B16 have demonstrated greater accuracy for classifying complex feature relationships as well as identifying minority classes. Specifically, for the squamous cell carcinoma (SCC) class (which was the least represented in the dataset), ViT-B16 achieved a recall of 87% and a precision of 72%. Despite promising results based on the proposed comparative framework, there are still a number of limitations. The comparative analysis was conducted using one publicly available dataset with no external cross-dataset validation, therefore limiting its generalizability to other clinical settings. Furthermore, the impact of staining artefacts (severe artefacts) as well as noise variation and imaging artefacts due to differences in laboratory settings were not systematically assessed. Therefore, there will be future studies that will explore cross-dataset validation, larger multi-centre datasets, analysis of computational efficiency, and hybrid CNN/transformers to further enhance robustness and clinical applicability in real-world scenarios.

VII. Conclusion

This study compared eight convolutional neural networks and a vision transformer model on a cervical cytology image dataset. The Vision Transformer (ViT-16) obtained the highest classification performance, while EfficientNet-B0 and DenseNet-121 also showed relatively strong results, including on minority classes such as SCC. Older models like AlexNet and NIN had more difficulty with class imbalance and deeper feature learning.

Overall, the findings suggest that models incorporating attention mechanisms or compound scaling provide advantages for medical image classification under imbalanced conditions. Future work may involve fine-tuning transformer models, exploring ensemble approaches, or integrating domain knowledge to improve classification reliability. For future directions, cross-dataset validation and the development of larger, multi-centre datasets will be pursued

A. Limitations

This study has several constraints. All models were trained from scratch without pretrained weights, which may affect generalization. The dataset was limited in size and imbalanced, especially for SCC, which may have influenced performance. In addition, results were obtained from a single dataset, and external validation would be useful for assessing robustness. Finally, computational efficiency was not the main focus, which could affect practical deployment in clinical environments.

Authors' Contribution

Dr. Mehreen Sirshar and Dr. Fakeeha Jaferi have supervised this work. Ibrahim Aljubayri and Abdulrahman Alahmadi prepared the figures. Omama Shakeel and Nayab Asim prepared the manuscript. Muhammad Zubair Khan reviewed the manuscript.

Non-financial Competiting interest. No, I declare that the authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this study.

Funding Declaration: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

This research is funded by the Deanship of Scientific Research, Islamic University of Madinah, Madinah, Saudi Arabia.

References

[1] Delgado Soler, D. (2022). Cervical cancer: a systematic review.

[2] G. Nirmala, P. P. Nayudu, A. R. Kumar, R. Sagar, "Automatic cervical cancer classification using adaptive vision transformer encoder with CNN for medical application," Pattern Recognition, vol. 160, p. 111201, 2025.

[3] P. Huang, X. Tan, C. Chen, X. Lv, Y. Li, "AF-SENet: Classification of cancer in cervical tissue pathological images based on fusing deep convolution features," Sensors, vol. 21, no. 1, p. 122, 2020.

[4] W. Liu, C. Li, N. Xu, T. Jiang, M. M. Rahaman, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, et al., "CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, Visual Transformer and Multilayer Perceptron," Pattern Recognition, vol. 130, p. 108829, 2022.

[5] J. Gangrade, R. Kuthiala, S. Gangrade, Y. P. Singh, S. Solanki, "A deep ensemble learning approach for squamous cell classification in cervical cancer," Scientific Reports, vol. 15, no. 1, p. 7266, 2025.

[6] Y. Karasu Benyes, E. C. Welch, A. Singhal, J. Ou, A. Tripathi, "A comparative analysis of deep learning models for automated cross-preparation diagnosis of multi-cell liquid Pap smear images," Diagnostics, vol. 12, no. 8, p. 1838, 2022.

[7] V. Sellamuthu Palanisamy, R. K. Athiappan, T. Nagalingam, "Pap smear based cervical cancer detection using residual neural networks deep learning architecture," Concurrency and Computation: Practice and Experience, vol. 34, no. 4, p. e6608, 2022.

[8] S. L. Tan, G. Selvachandran, W. Ding, R. Paramesran, K. Kotecha, "Cervical cancer classification from Pap smear images using deep convolutional neural network models," Interdisciplinary Sciences: Computational Life Sciences, vol. 16, no. 1, pp. 16–38, 2024.

[9] D. C. R. Novitasari, P. Wulandari, D. Z. Haq, "Cervical cancer diagnosis system using convolutional neural network ResidualNet," International Journal of Computing, vol. 21, no. 1, pp. 61–68, 2022.

[10] S. Cheng, S. Liu, J. Yu, G. Rao, Y. Xiao, W. Han, W. Zhu, X. Lv, N. Li, J. Cai, et al., "Robust whole slide image analysis for cervical cancer screening using deep learning," Nature Communications, vol. 12, no. 1, p. 5639, 2021.

[11] H. Kaur, R. Sharma, J. Kaur, "Comparison of deep transfer learning models for classification of cervical cancer from Pap smear images," Scientific Reports, vol. 15, no. 1, p. 3945, 2025.

[12] S. Alsubai, A. Alqahtani, M. Sha, A. Almadhor, S. Abbas, H. Mughal, M. Gregus, "Privacy preserved cervical cancer detection using convolutional neural networks applied to Pap smear images," Computational and Mathematical Methods in Medicine, vol. 2023, no. 1, p. 9676206, 2023.

[13] F. Paucar, C. Bojorque, I. Reyes-Chac´on, P. Vizcaino-Imacan˜a, M. MorochoCayamcela, "Towards accurate cervical cancer detection: Leveraging two-stage CNNs for Pap smear analysis," Proc. Int. Conf., pp. 219–227, 2024.

[14] C. Zhao, R. Shuai, L. Ma, W. Liu, M. Wu, "Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT," Multimedia Tools and Applications, vol. 81, no. 17, pp. 24265–24300, 2022.

[15] E. Lima, J. Smith, and R. Kumar, "A novel approach to medical image classification," IEEE Transactions on Medical Imaging, vol. 42, no. 3, pp. 456–465, 2024.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021.

[17] AlexNet, https://cdn.techscience.cn/ueditor/files/iasc/TSPIASC-36-1/TSPIASC30125/T-SPIASC30125.pdf, accessed: 2025-04-14.

[18] VGG, https://media.geeksforgeeks.org/wp-content/uploads/20200219152327/convlayers-vgg16-1024x450.jpg, accessed: 2025-04-14.

[19] GoogLeNet, https://miro.medium.com/v2/resize:fit:4800/format:webp/0*W5DczbPesqqmEWO.jpg, accessed: 2025-04-14.

[20] Network in Network https://www.researchgate.net/figure/fig3/AS:915542533750784159529 3758115

[21] /The-structure-of-the-NIN-model-used-in-our-proposed-method-The-definitionsof-all.png, accessed: 2025-04-14.

[22] ResNet, https://towardsdatascience.com/wp-content/uploads/2022/08/0tH9evuOFqk8F41FG2048x660.png, accessed: 2025-04-14.

[23] DenseNet, https://media.springernature.com/full/springer-static/image/art, accessed: 2025-04-14.

[24] E. Hussain, "Liquid based cytology Pap smear images for multi-class diagnosis of cervical cancer," Mendeley Data, 2019.

[25] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, vol. 25, 2012.

[26] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning," Image Recognition, vol. 7, no. 4, pp. 327–336, 2015.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

[28] G. Huang, Z. Liu, K. Q. Weinberger, "Densely connected convolutional networks," CoRR, abs/1608.06993, 2016. [ONMILine]. Available: http://arxiv.org/abs/1608.06993

[29] M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, PMLR, pp. 6105–6114, 2019.

[30] M. Lin, Q. Chen, S. Yan, "Network in network," arXiv preprint arXiv:1409.1556, 2014.

[31] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[32] Kumar, G. V., Bellary, M. I., Reddy, T. B. (2022). Prostate cancer classification with MRI using Taylor-Bird squirrel optimization based deep recurrent neural network. The Imaging Science Journal, 70(4), 214-227.

[33] Kumar, A., Ahmad, F., Alam, B. (2025). Hybrid bio-inspired computing in medical image data analysis: A review. Intelligent Decision Technologies, 19(1), 473-488.

[34] Bohmrah, M. K., Kaur, H. (2025). Advanced hybridization and optimization of DNNs for medical imaging: A survey on disease detection techniques. Artificial Intelligence Review, 58(4), 122.

Methodology for Identifying Soiling on PV Panels Using RGB Images and Deep Learning

Katerina Gabrovska-Evstatieva[1], Tsvetelina Kaneva[2], Irena Valova[3], Dimitar Trifonov[4], Nikolay Valov[5], Ventsislav Keseev[6], Nicolay Mihailov[7], Boris Evstatiev[8]
Faculty of Natural Science and Education, University of Ruse Angel Kanchev, Ruse 7004, Bulgaria [1]
Faculty of Electrical Engineering, Electronics and Automation,
University of Ruse Angel Kanchev, Ruse 7004, Bulgaria[2, 3, 4, 5, 6, 7, 8]

***Abstract*—Soiling is one of the major factors that can affect the performance of PV installations in many regions of the world. On the other hand, the variety of PV models, types of soiling, and climate conditions makes it very difficult to create a universal model. To address this problem, this study presents a methodology for the identification of soiling on PV panels via semantic segmentation, which can support the decision-making process in terms of surface cleaning and automation of the process. The methodology includes data collection, preparation of training and testing data, training of models, and application of the optimal one. Next, the methodology is demonstrated using a small dataset of clean and dirty PV panels, three neural network architectures (DeepLab v3, U-Net, and PSPNet), and two backbone models (ResNet34 and ResNet50). The obtained results show the feasibility of the methodology and allow highlighting the DeepLab v3 model with a ResNet34 backbone as the best-performing algorithm for identifying pigeon droppings. The second-best combination is the U-Net + ResNet34, which showed good efficiency for identifying smaller dirty areas. The proposed methodology could be useful for operators of large-scale photovoltaic installations by supporting the decision-making process when it comes to the timely cleaning of specific areas for performance improvement and lower costs.**

***Keywords*—PV soiling; semantic segmentation; pixel-based classification; neural network; RGB images**

I. Introduction

In times of energy crisis and a changing climate, the importance of renewable energy sources (RES) increases significantly. Out of them, photovoltaic (PV) systems are dominating the renewables market due to their versatility and low-maintenance requirements. Nevertheless, numerous factors exist that influence PV installations' performance and, as a result, open many opportunities for their optimization. Two of these factors are soiling and shading of the photovoltaic surface, which could have a great impact on a PV installation's performance [1,2]. They are known to lead to various consequences, such as reduced energy yield [3-5], increased panel degradation [5,6], hotspots [7], fire hazard [8,9], etc. In most situations, shading is a factor that can be acknowledged during design time, but there are limited options for its mitigation.

Soiling, on the other hand, could be effectively mitigated during exploitation; however, there are a number of factors to consider. According to [10], soiling accumulation can reduce energy production by up to 20%, depending on the environmental conditions. In [11] was obtained that agricultural soiling could reduce light transmission by up to 40% and the power output by up to 10%. Similar results about bird droppings showed that the power losses could reach a maximal value between 10% and 23%, depending on the month of the year and the PV installation location [12].

To mitigate the impact of soiling, different approaches for cleaning the photovoltaic surfaces exist. The most basic one is manual cleaning, which is known to be both expensive and time-consuming [13]. Furthermore, it requires water that is free from minerals to prevent the creation of spots and scale build-up. Other studies have suggested using the thrust of a drone to clean dust from the PV surface [14] or different types of robots moving over the PV surface [15,16]. In [17,18], a mechanism for self-cleaning panels' surface by removing dust particles using spherical water droplets is proposed, where a sliding or rolling motion is imparted. In all cases, the price is a major factor, and therefore, it is important to properly schedule the cleaning cycles [19]. The importance of properly deciding on when to clean the PV surface increases even more for building-integrated photovoltaics, where conventional cleaning approaches are rarely applicable [20].

Another problem is the type of soiling. Most of the above-mentioned studies propose solutions for removing dust. However, periodic cleaning, especially if a waterless method is used, is not always a suitable solution for bird droppings, which are especially common in the urban environment. All of this shows that the identification of bird droppings is a major factor for the optimization of the PV cleaning process.

Numerous studies have proposed a wide range of approaches for identifying soiling on PV surfaces based on image processing and classification [21], RGB image analysis [22], and pixel- and object-based identification. In [10], a methodology for the classification of RGB images of PV surfaces as either clean or dirty was proposed. The images were preliminarily processed using red channel doubling and conversion to multichannel, which allowed the dirty parts to be highlighted. Different machine learning algorithms were tested, and the best results were achieved by an artificial neural network, reaching an F1 score of 91.3%. In another study [23], deep learning was used to classify RGB images of PV modules into one of the 5 classes: very unclean, slightly unclean, medium clean, slightly clean, and clean. The optimal results

were achieved by the UTran neural network architecture with a ResNet50 backbone, reaching an accuracy of 96.41%.

In [24], a pixel-based classification approach was used to identify hotspots on PV surfaces by classifying the pixels of IR images using machine learning algorithms. The output of the models is a mask that identifies potential problems. A different approach was used in [25] with images in the visible spectrum, aimed at teaching a robot to identify and clean soiled surfaces automatically. The study proposed a threshold value used on the red channel of the RGB image in order to create the cleaning masks. In [26], several methods for semantic segmentation were used to generate soiling masks on PV surfaces. The algorithms used include Random Forest, XGBoost, LightGBM, and the U-Net neural network architecture. The neural network achieved the highest performance, reaching an accuracy and F1 score of 96.82% and 93.37%, respectively.

Other studies investigated the object-based approach for identifying soiling on PV panels. In [27], RGB images were used with the SDS-YOLO v5 architecture. The model achieved F1 scores of 0.700 and 0.781 for the bird droppings and dust classes, respectively. A similar approach was used in [28], where the YOLO v8 model's soiling identification efficiency exceeded 90%. In [29], different classification models were compared to identify PV performance degradation caused by dust and dirt accumulation. The results demonstrated the potential of data-driven approaches to improve the monitoring and maintenance of photovoltaic systems and to optimize energy efficiency.

A completely different approach was used in [30], where real-world meteorological and production data were used to develop and validate a predictive algorithm based on Stacked Ensemble Learning (SEL) and Decision Trees (DT). It achieved high accuracy in predicting soiling and panel performance, which shows that such models can help optimize maintenance and improve the energy efficiency of photovoltaic systems.

The performed analysis shows that numerous studies have proposed different approaches for the identification of soiling on PV surfaces with AI algorithms, with the object-based ones commonly having a lower accuracy compared to the pixel- and classification-based ones. Some of them achieved quite impressive results, yet their accuracy generally becomes lower when used on different datasets, which is normal for machine learning. The various PV technologies have different appearances in terms of color, cell dimensions, etc., which is a major limiting factor for the models to be more universal. Furthermore, the soiling factors also vary in different geographic locations. This means that in order to achieve optimal results, the models should be trained with a dataset that was obtained for a specific PV facility. Most studies concentrate on the modelling part, rather than the methodological aspects, which are more important as they allow for obtaining improved results for the local specifics. Therefore, this could be considered a research gap.

Considering the above-mentioned, this study aims to propose a methodology for assessing the soiling of PV panels, which allows for training high-accuracy deep learning models capable of supporting the decision-making process in terms of organizing the cleaning routines. The feasibility of the proposed methodology is demonstrated using a small dataset with PV modules under the impact of pigeon droppings, obtained from the region of Ruse, Bulgaria. The study also evaluates the performance of different deep learning architectures and backbone models for identifying bird droppings with different impact areas.

II. Materials and Methods

A. Methodology of the Study

The proposed methodology for assessing the condition of a PV surface is summarized in Fig. 1. It can be described in 4 steps as follows:

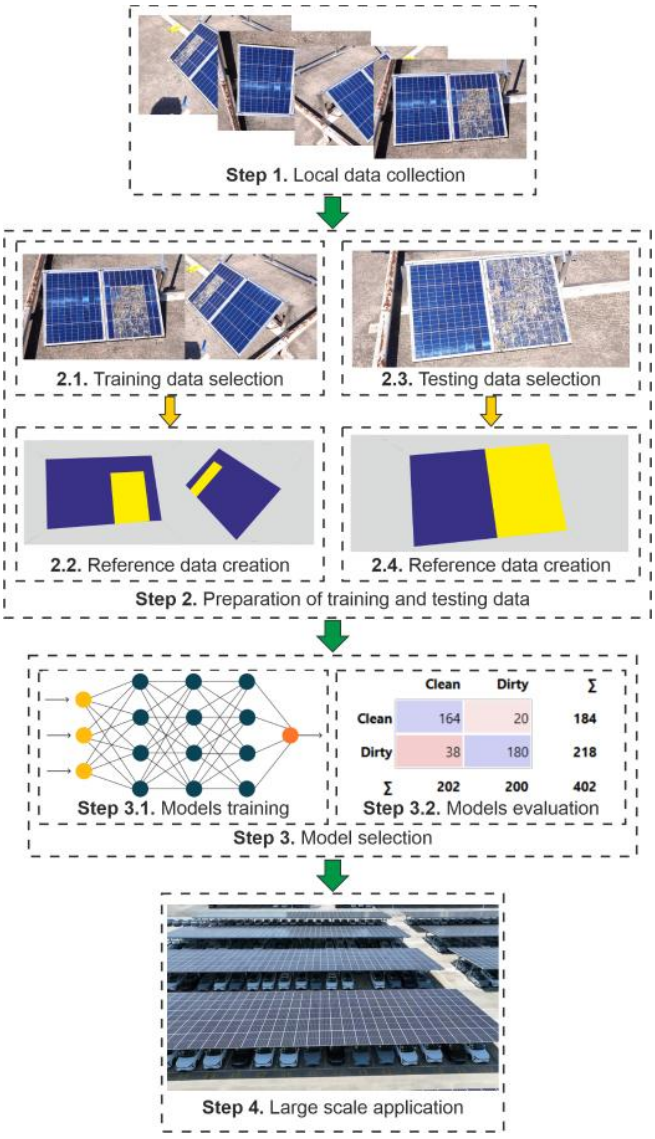


Fig. 1. Methodology for identification of soiling on a PV surface.

Step 1. Local data collection

This step includes the collection of images from the specific PV installation that is to be evaluated. The data must be obtained for the specific facility for several reasons:

- The way the panels look depends on their technology, weather conditions, etc., which is always site-specific;
- The potential soiling is also site-specific as it depends on the local soil types, the type of birds, etc.

Furthermore, when collecting the local dataset, it is recommended that the following requirements be met:

- It should represent all potential types of soiling on the PV surface;
- It should include surfaces of well-cleaned panels;
- It should include the different background surfaces available at the site (grass, concrete, etc.).

Step 2. Preparation of training and testing data

Next, training and testing data should be prepared, which is implemented in two similar processes.

1) Training data selection: In this phase, several images are selected from the local dataset, which represent all major situations, such as clean panels, panels with different types of soiling, and images with different types of surrounding backgrounds. It is recommended that the different situations are more or less equally represented in order to improve the performance of the models. According to the proposed methodology, the selected images should be merged into a single larger image.

2) Reference data creation: Next, using the compiled training image, reference data should be created. In this study, the following classes are used:

- Clean PV surface ("Clean") – an area of the PV surface, which is in good condition and no soiling can be observed;
- Dirty PV surface ("Dirty") – an area of the PV surface, on which soiling can be observed;
- Background ("Back") – everything else that is not part of the PV surface (whether clean or dirty) is considered a surrounding or background area.

If necessary, the "Dirty PV surface" class could be divided into subclasses, representing different types of soiling. However, in this study, we have only investigated soiling caused by bird droppings.

In this phase, the pixels of the compiled training image are manually classified into one of the selected categories. An important recommendation is that if the operator cannot clearly identify the class of a certain area of the image, the corresponding pixels should remain uncategorized.

3) Testing data selection: In this phase, other images should be selected for testing purposes that are different from the training ones. They should also adequately represent all common situations. Similarly, the images could be merged into a single testing image for easier processing and analysis.

4) Reference data creation: This phase is similar to 2.2 and includes classifying the pixels of the testing image into the adopted classes.

Step 3. Model selection

This step includes two phases – training models and evaluating their performance, aimed at identifying the optimal ones.

1) Models' training: In this study, three deep learning architectures are selected:

- U-Net- is widely used for pixel-based classification and has shown good performance for dust detection in photovoltaic panels [31]. It is also known for achieving high accuracy with fine details, being extremely efficient when working with small data sets, and typically requires less training time and hardware resources [32]. Despite being lighter and faster, it often performs relatively well compared to other models [33];
- DeepLab v3- is another semantic segmentation model, which is known to return good results with the identification of photovoltaic installations [34,35], surface soiling [36], hot-spot defects [37], etc. It usually returns better accuracy at different object scales and more complex images, but can be significantly slower to train and requires a larger amount of data and hardware resources [38];
- Pyramid Scene Parsing Network (PSPNet)- has also shown good results with PV panel identification [35]. It is known for better distinguishing objects based on analysis of the surrounding environment, but can be significantly slower to train and requires a larger amount of data and hardware resources, although some works report that it is the most efficient in terms of execution time [39].

Furthermore, for each one of them, two backbone models have been used: ResNet-34 and ResNet-50, thus making a total of 6 combinations. The difference between ResNet-34 and ResNet-50 lies mainly in their depth and architectural complexity. ResNet-34 is a lighter backbone model, requiring less training data for simpler tasks, training faster, and taking up less memory on the graphics processing unit. ResNet-50, on the other hand, is more complex and is better at recognizing fine textures and complex objects, but requires more data to avoid overfitting. Furthermore, it requires more computing power and memory, and training takes longer.

This phase of the methodology includes training the models using the training data, prepared in phases 2.1 and 2.2 of the methodology.

2) Models' evaluation: Next, each of the trained models is applied to the testing data prepared in phases 2.3 and 2.4, and a classification image is created. Thereafter, the performance of each model is evaluated against the reference data using the following metrics:

- Accuracy – estimates the proportion of total correct predictions:

$$Acc = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative} \quad (1);$$

- Precision – estimates the share of true positives out of all positively identified pixels:

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \quad (2);$$

- Recall – estimates the share of the correctly identified pixels:

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad (3);$$

- F1 score – averages the precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4);$$

- Cohen's Kappa – evaluates the level of agreement between classified and reference data, with 0 corresponding to no agreement at all (all pixels were classified incorrectly) and 1 corresponding to perfect agreement (all pixels were classified correctly) [40].

The above metrics are used to select the best-performing model(s).

Step 4. Large-scale application

Once the best-performing model is selected, it can be used for large-scale identification of soiling at the specific photovoltaic facility as an instrument for automating the decision-making and cleaning processes.

B. Means of the Investigation

In this study, the software ArcGIS Pro v. 3.6.1 is used, developed by Esri Inc. Even though this tool was created as a GIS analysis instrument, it is also applicable to images that are not mapped to a specific coordinate system.

ArcGIS Pro provides a wide range of object- and pixel-based machine learning and deep learning algorithms, integrated tools for the creation of reference data, comparison between the modeled and reference data, generation of confusion matrices, estimation of different measures, etc. Furthermore, all of the above is implemented in a friendly graphical user interface, i.e., no specialized programming knowledge and skills are required. This means that with appropriate instructions, the ArcGIS Pro tool could be used by operators of photovoltaic installations without the need for IT-specific qualifications.

C. Organization of the Experiment

The experiment in this study was performed at the site of PV Park Kanev, located in the University of Ruse "Angel Kanchev", Ruse, Bulgaria, geographic coordinates 43.85344121491259, 25.969174955358763 (Fig. 2). Two 50W polycrystalline PV panels with model CL-SM50P were used in the experiment. One of the two panels has been placed on the floor of a dovecote for two weeks. This allowed its surface to be densely covered with pigeon droppings, which are one of the main sources of PV soiling in the Bulgarian urban landscape.



Fig. 2. Location of the experimental site.

III. Results and Discussion

In this section, the study aims to demonstrate the feasibility of the proposed methodology using a limited amount of input data. The clean and dirty PV panels were installed at the experimental site using a specially created mounting structure, and numerous images of them were made from different angles. Next, different areas of the soiled PV surface were gradually cleaned, and additional photos were made at each stage (Fig. 3). The data collection was finalized when both PV surfaces were clean.



Fig. 3. Photos of the two PV panels with different levels of dirtiness.

Next, according to Step 2 of the methodology, four images were selected, which represent different aspects of the experimental situation, such as:

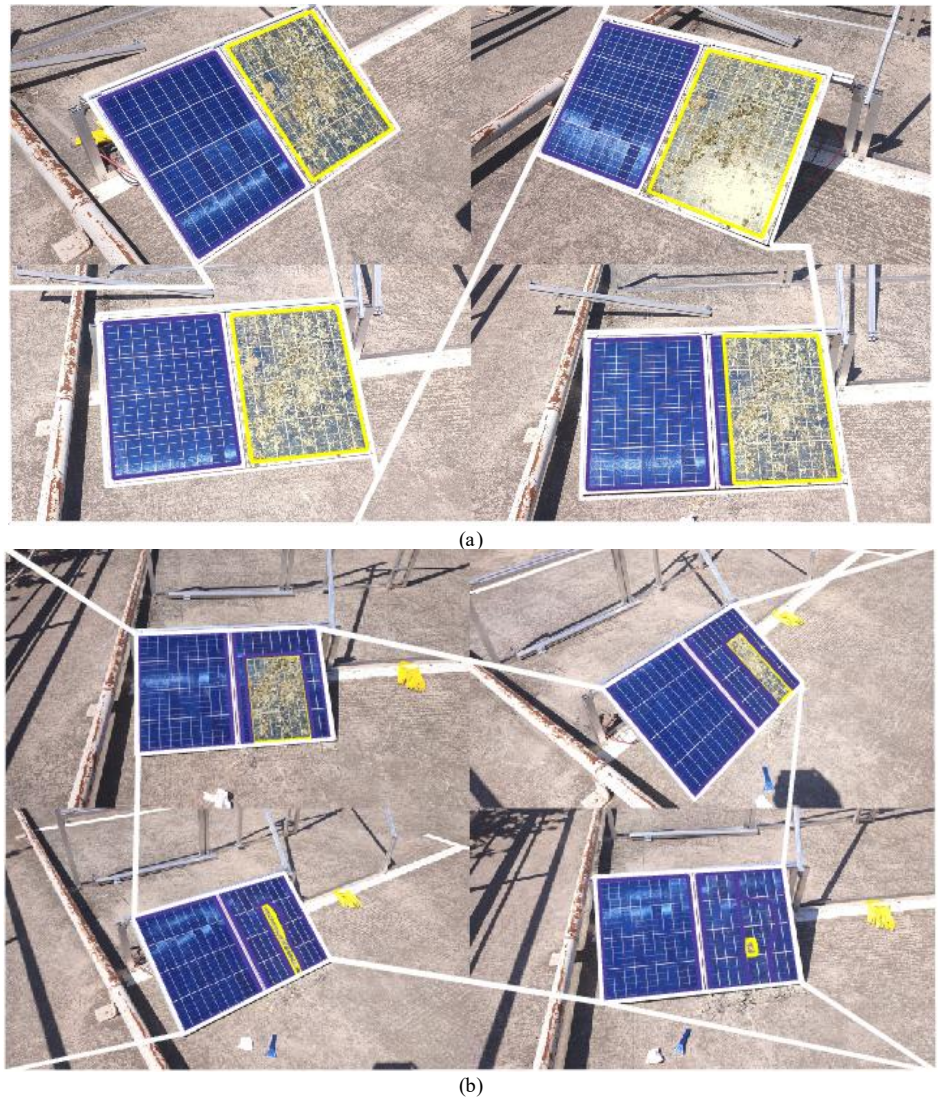
- The PV modules are photographed from different angles, which represent the surfaces differently affected by the solar rays;

- The surrounding area is observed from different angles, which influences its view, the impact of shadows, etc.

The four images were merged into a single one using Microsoft's MS Paint tool. Thereafter:

- The prepared training image was imported into ArcGIS Pro v. 3.6.1;
- Reference areas were created for the three classes (Clean, Dirty, and Background) with the Image Classification tool using polygons (Fig. 4a).

Finally, the created reference data was exported in the "Classified tiles" metadata format with X/Y tile size 256 and X/Y Stride 128. Similarly, four testing images were selected, merged into a single image, and reference data was created for it (Fig. 4b).



(a)

(b)

Fig. 4. The selected training (a) and testing (b) images with the marked reference polygons.

Following Step 3 of the methodology, 6 models were trained using the three NN architectures and the two backbone models on an HP Pavilion laptop with an NVidia GeForce MX350 GPU. 20 epochs were used for the training, 10% of the input data was used for validation, the optimal learning rate was automatically extracted from the learning curve, and all other parameters were kept at their default values. The obtained accuracies and F1 scores for the six models are summarized in Table I. It can be seen that if accuracy and F1 score are used as the main indicators, the DeepLabv3+ResNet50 combination is the worst-performing model. The differences between the measures of the other 5 models are 1-2 thousandths, and therefore could be considered insignificant. In general, the U-Net and PSPNet-based models achieve the highest performance.

On the other hand, when the Kappa coefficient is used, the PSPNet-based models have the lowest performance, while the U-Net+ResNet34 combination is the second best. The DeepLabv3+ResNet34 achieved the best performance, reaching a Kappa score of 0.996. It is important to mention that Cohen's Kappa is considered a more stable indicator than Accuracy because it excludes the chance of hitting the result when some data predominates [41].

TABLE I. SUMMARY OF THE SIX MODELS' TRAINING SCORES

№	Model	Accuracy	F1 Score	Kappa
1	DeepLab v3 + ResNet34	0.982	0.987	0.996
2	DeepLab v3 + ResNet50	0.975	0.980	0.985
3	U-Net + ResNet34	0.983	0.987	0.988
4	U-Net + ResNet50	0.984	0.988	0.980
5	PSPNet + ResNet34	0.983	0.987	0.954
6	PSPNet + ResNet50	0.983	0.988	0.962

The conclusion from the measures analysis is that the DeepLab v3+ResNet34 model, even though requiring more resources and time, could perform slightly better, especially in situations such as the current one, when comparatively little data is available. U-Net+ResNet34 is expected to perform about the same, but with less required time and resources. However, according to [42], DeepLabV3+ outperforms U-Net in terms of both speed and accuracy.

Next, the six models were applied to the testing image, and the obtained classification images are presented in Fig. 5. At first glance, it can be seen that all models perform very well, though some false positives can be observed. Furthermore, it can be observed that larger dirty areas are well identified by all models, while the results with smaller and thinner ones are more variable, and some false positives and false negatives can be observed.

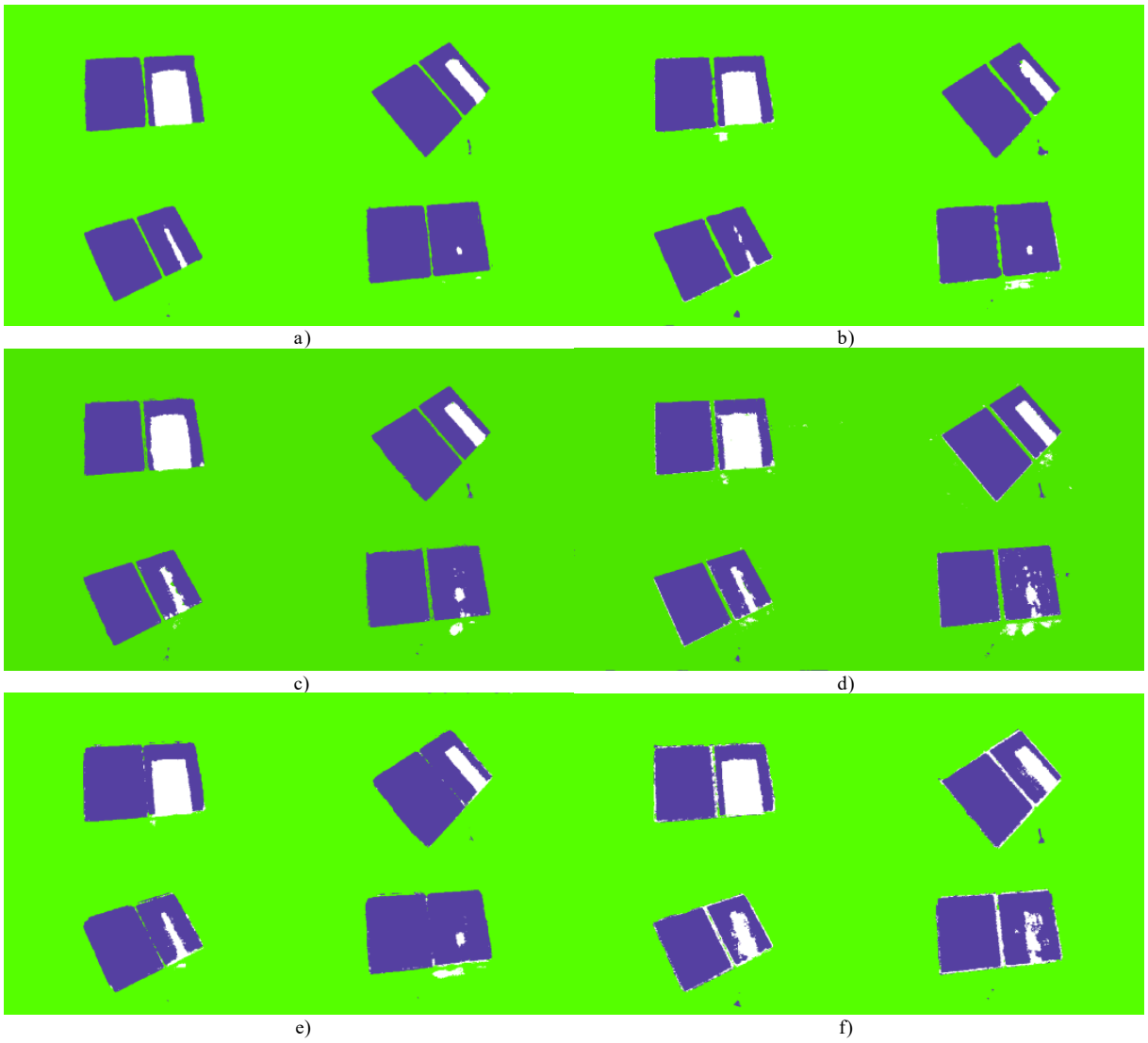


Fig. 5. Classification results for the testing image done with: DeepLab v3 + ResNet34 (a); DeepLab v3 + ResNet50 (b); U-Net + ResNet34 (c); U-Net + ResNet50 (d); PSPNet + ResNet34 (e); PSPNet + ResNet50 (f).

To get a better understanding of the models' performances, they were assessed using the "Accuracy assessment" tool in ArcGIS, 10,000 randomly selected pixels, and the Stratified random strategy was used on the reference testing data to generate the confusion matrices for the six situations.

The results for the DeepLab v3 + ResNet34 model are presented in Table II. It can be seen that the model has almost perfect performance, with an average F1 score of 0.999. The lowest precision and recall are obtained for the Dirty class, reaching values of 0.974 and 0.993, respectively, which indicates the classification of some false positives.

TABLE II. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE DEEPLAB V3 + RESNET34 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1332	4	5	1341	0.999	0.993	0.996
	Dirty	0	149	1	150	0.974	0.993	0.983
	Back	1	0	8508	8509	0.999	1.000	1.000
	Total	**1333**	**153**	**8514**	**10000**	**0.999**	**0.999**	**0.999**

Table III presents the performance of the DeepLab v3 + ResNet50 model. Its metrics are slightly lower for the clean and background classes, as for the dirty class, the precision and recall have dropped significantly, reaching values of 0.902 and 0.890, resulting in an F1 score of 0.896.

TABLE III. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE DEEPLAB V3 + RESNET50 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1330	14	6	1350	0.999	0.985	0.992
	Dirty	0	138	17	155	0.902	0.890	0.896
	Back	1	1	8490	8492	0.997	1.000	0.999
	Total	1331	153	8513	9997	0.996	0.996	0.996

Next, in Table IV and Table V, the measures for the two U-Net models are shown. Unlike the DeepLab models, these achieved higher precision for the dirty class, reaching up to 1.000 with the ResNet50 backbone; however, the recall is lower, which indicates an increased share of false positives. These results confirm the classification images in Fig. 5, which also show the identification of many false-positive "dirty" pixels.

TABLE IV. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE U-NET + RESNET34 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1320	0	10	1330	0.990	0.992	0.991
	Dirty	13	154	7	174	0.994	0.885	0.936
	Back	0	1	8495	8496	0.998	1.000	0.999
	Total	1333	155	8512	10000	0.997	0.997	0.997

TABLE V. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE U-NET + RESNET50 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1311	0	3	1314	0.986	0.998	0.992
	Dirty	19	153	30	202	1.000	0.757	0.862
	Back	0	0	8481	8481	0.996	1.000	0.998
	Total	1330	153	8514	9997	0.995	0.995	0.994

Finally, in Table VI and Table VII, the results for the PSPNet models are summarized. Similarly to the U-Net models, the dirty class precision is 1.000, which indicates that all dirty pixels were correctly identified. However, the recall decreases even more, reaching 0.748 and 0.698 with the ResNet34 and ResNet50 backbones, respectively. Once again, this indicates that these models return too many false positives when it comes to the dirty areas of the PV panel, which might lead to misleading results and have a significant impact on the decision-making process.

TABLE VI. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE PSNET + RESNET34 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1325	0	71	1396	0.993	0.949	0.971
	Dirty	9	154	43	206	1.000	0.748	0.856
	Back	0	0	8400	8400	0.987	1.000	0.993
	Total	1334	154	8514	10002	0.988	0.988	0.987

TABLE VII. CONFUSION MATRIX AND MEASURES FOR THE TESTING DATA WITH THE PSNET + RESNET50 MODEL

		Predicted				Metrics		
		Clean	***Dirty***	***Back***	***Total***	***Prec.***	***Rec.***	***F1***
Actual	***Clean***	1300	0	33	1333	0.975	0.975	0.975
	Dirty	33	155	34	222	1.000	0.698	0.822
	Back	0	0	8447	8447	0.992	1.000	0.996
	Total	1333	155	8514	10002	0.990	0.990	0.989

The obtained results show that the DeepLab v3 model with a ResNet34 backbone achieved the best performance, followed by the U-Net + ResNet34 one, for the following reasons:

- All models achieved more or less similar performance for the clean and background classes, with their precisions/recalls varying between 0.975 and 1.000;
- Most models achieved great precision for the dirty class, varying between 0.974 and 1.000 (the only exception is the DeepLab v3 + ResNet50 model, which achieved a precision of 0.902);
- The DeepLab v3 + ResNet34 model achieved the highest recall for the dirty class, reaching a value of 0.993, while this measure for all other models is significantly lower, ranging between 0.698 and 0.890.

These results also correspond with those obtained using the Kappa coefficient during the training phase. They confirm that Cohen's Kappa is a more reliable measure and could be used as the main indicator for selecting the optimal model.

IV. CONCLUSION

The current study presents a methodology for the identification of soiling on PV panels, which is optimized for providing reliable site-specific results. It can be used as a supporting instrument in the decision-making process of cleaning large photovoltaic installations. The methodology includes data collection, preparation of training and testing

data, training of models, and identification of the optimal one, which can then be applied on a large scale.

Next, the feasibility of the methodology is demonstrated by using limited experimentally obtained data from the PV park Kanev, Ruse, Bulgaria. An image dataset was created with clean and dirty PV panels under the impact of pigeon droppings. Training and testing images were selected, and reference data were created for them using three classes: "clean", "dirty", and "background". Three deep learning architectures were used: DeepLab v3, U-Net, and PSPNet. Furthermore, each one of them was trained with the ResNet34 and ResNet50 backbones. All trained models achieved more or less similar metrics and very good performance.

However, after they were applied to a testing dataset, the DeepLab v3 model with a ResNet34 backbone showed the highest performance, expressed in the best ratio between precision and recall regarding the "dirty" class. The second-best combination is the U-Net + ResNet34, which proved to be extremely efficient when working with small datasets. Both models could be applied in practice for the identification of soiling areas with relatively small datasets specific to the photovoltaic site. The study results also showed that Cohen's Kappa could be used as the main indicator for selecting the optimal model of the methodology, which is an important observation for the future application of the methodology.

The methodology proposed in this study could be useful for operators of large photovoltaic installations by supporting the decision-making process when creating the cleaning schedule. Even though the obtained results are quite promising, they have some limitations. They were not tested with real data, obtained from drones or satellites, which is an important goal for future research. Furthermore, it is worth investigating the possibility of increasing the performance of all algorithms by using a secondary model for limiting the analyzed area only to the area of PV modules, thus allowing the removal of the "background" class.

ACKNOWLEDGMENT

This study is financed by the European Union—NextGenerationEU through the National Recovery and Resilience Plan of the Republic of Bulgaria, project no. BGRRP-2.013-0001.

REFERENCES

[1] J. A. Tsanakas et al., "Assessment of Soiling Dynamics and Cleaning Efficiency for Photovoltaic Modules Under Different Dust Environments," Solar RRL, vol. 10, no. 6, Dec. 2025, doi: 10.1002/solr.202500792.

[2] R. K. Pachauri et al., "Impact of Partial Shading on Various PV Array Configurations and Different Modeling Approaches: A Comprehensive Review," in IEEE Access, vol. 8, pp. 181375-181403, 2020, doi: 10.1109/ACCESS.2020.3028473.

[3] Y. Mhanni and Y. Lagmich, "Maximizing Solar Panel Efficiency in Partial Shade: The Improved POA Solution for MPPT," International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, 2024, doi: 10.14569/ijacsa.2024.0150397.

[4] M. R. Maghami, H. Hizam, C. Gomes, M. A. Radzi, M. I. Rezadad, and S. Hajighorbani, "Power loss due to soiling on solar panel: A review," Renewable and Sustainable Energy Reviews, vol. 59, pp. 1307–1316, Jun. 2016, doi: 10.1016/j.rser.2016.01.044.

[5] M. J. Alshareef, "A Comprehensive Review of the Soiling Effects on PV Module Performance," in IEEE Access, vol. 11, pp. 134623-134651, 2023, doi: 10.1109/ACCESS.2023.3337204.

[6] K. G. Gabrovska-Evstatieva, D. T. Trifonov, and B. I. Evstatiev, "A Review of the Key Factors Influencing the Performance of Photovoltaic Installations in an Urban Environment," Electricity, vol. 6, no. 2, p. 23, May 2025, doi: 10.3390/electricity6020023.

[7] M. Bressan, A. Gutierrez, L. Garcia Gutierrez, and C. Alonso, "Development of a real-time hot-spot prevention using an emulator of partially shaded PV systems," Renewable Energy, vol. 127, pp. 334–343, Nov. 2018, doi: 10.1016/j.renene.2018.04.045.

[8] Z. Wu, Y. Hu, J. X. Wen, F. Zhou, and X. Ye, "A Review for Solar Panel Fire Accident Prevention in Large-Scale PV Applications," in IEEE Access, vol. 8, pp. 132466-132480, 2020, doi: 10.1109/ACCESS.2020.3010212.

[9] A. Lysyi, A. Sachenko, P. Radiuk, M. Lysyi, and O. Savenko, "Enhanced fire hazard detection in solar power plants: an integrated UAV, AI, and SCADA-based approach," RADIOELECTRONIC AND COMPUTER SYSTEMS, vol. 2025, no. 2, pp. 99–117, May 2025, doi: 10.32620/reks.2025.2.06.

[10] B. I. Evstatiev, D. T. Trifonov, K. G. Gabrovska-Evstatieva, N. P. Valov, and N. P. Mihailov, "PV Module Soiling Detection Using Visible Spectrum Imaging and Machine Learning," Energies, vol. 17, no. 20, pp. 5238–5238, Oct. 2024, doi: 10.3390/en17205238.

[11] P. Gebhardt, P. Katouli, T. Kaltenbach, T. Bretzel, L.-M. Bieber, I. Haedrich, "EVALUATING THE IMPACT OF SOILING ON AGRIVOLTAIC SYSTEMS," European Photovoltaic Solar Energy Conference and Exhibition 2025, 2025, available online: https://publica-rest.fraunhofer.de/server/api/core/bitstreams/82384218-ec05-4370-9fe7-7b6e292e0b9c/content.

[12] A. K. Sisodia, "Effect of bird dropping deposition on photovoltaic (PV) module performance at different sites," International Journal of Science and Research Archive, vol. 13, no. 2, pp. 3569–3573, Dec. 2024, doi: 10.30574/ijsra.2024.13.2.2566.

[13] S. Gochhait, R. Asodiya, T. Hasanmani, V. Patin, and O. Maslova, "Application of IoT: A Study on Automated Solar Panel Cleaning System," 2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, 2022, pp. 1-4, doi: 10.1109/ICECIE55199.2022.10000375.

[14] S. Rehman, M. A. Mohandes, A. E. Hussein, L. M. Alhems, and A. Al-Shaikhi, "Cleaning of Photovoltaic Panels Utilizing the Downward Thrust of a Drone," Energies, vol. 15, no. 21, p. 8159, Nov. 2022, doi: 10.3390/en15218159.

[15] S. A. O. Sanya, M. Houngbedji, M. F. O. Sanya, and D. C. Akowanou, "Design of a Robotic System for Cleaning Solar Panels in Benin," Open Journal of Applied Sciences, vol. 15, no. 04, pp. 962–975, 2025, doi: 10.4236/ojapps.2025.154065.

[16] G. Ashtaputre and A. Bhoi, "Artificial Intelligence Based Solar Panel Cleaning Robot," vol. 14, pp. 12–15, 2019, doi: 10.9790/2834-1403021215.

[17] X. Gong and S. He, "Highly Durable Superhydrophobic Polydimethylsiloxane/Silica Nanocomposite Surfaces with Good Self-Cleaning Ability," ACS Omega, vol. 5, no. 8, pp. 4100–4108, Feb. 2020, doi: 10.1021/acsomega.9b03775.

[18] S. Maharjan et al, "Self-cleaning hydrophobic nanocoating on glass: A scalable manufacturing process," Materials Chemistry and Physics, vol. 239, p. 122000, Jan. 2020, doi: 10.1016/j.matchemphys.2019.122000.

[19] M. Rashid, M. Yousif, Z. Rashid, A. Muhammad, M. Altaf, and A. Mustafa, "Effect of Dust Accumulation on the Performance of Photovoltaic Modules for Different Climate Regions," Heliyon, vol. 9, no. 12, pp. e23069–e23069, Dec. 2023, doi: 10.1016/j.heliyon.2023.e23069.

[20] S. Umar et al., "A building integrated solar PV surface-cleaning setup to optimize the electricity output of PV modules in a polluted atmosphere," Renewable Energy, vol. 216, pp. 119122–119122, Aug. 2023, doi: 10.1016/j.renene.2023.119122.

[21] U. Naeem, K. Chadda, S. Vahaji, J. Ahmad, X. Li, and E. Asadi, "Aerial Imaging-Based Soiling Detection System for Solar Photovoltaic Panel

Cleanliness Inspection," Sensors, vol. 25, no. 3, pp. 738–738, Jan. 2025, doi: 10.3390/s25030738.

[22] R. Cavieres, R. Barraza, D. Estay, J. Bilbao, and P. Valdivia-Lefort, "Automatic soiling and partial shading assessment on PV modules through RGB images analysis," Applied Energy, vol. 306, p. 117964, Jan. 2022, doi: 10.1016/j.apenergy.2021.117964.

[23] M. Li and Y. Wang, "Deep Learning for Dust Accumulation Analysis on Desert Solar Panels: A CNN-Transformer Approach," in IEEE Access, vol. 13, pp. 69857-69872, 2025, doi: 10.1109/ACCESS.2025.3561363.

[24] M. Abdelsattar, A. AbdelMoety, and A. Emad-Eldeen, "Applying Image Processing and Computer Vision for Damage Detection in Photovoltaic Panels," Mansoura Engineering Journal, vol. 50, no. 2, Mar. 2025, doi: 10.58491/2735-4202.3263.

[25] T. Pivem, F. de Oliveira de Araujo, L. de Oliveira de Araujo, and G. S. de Oliveira, "Application of a Computer Vision Method for Soiling Recognition in Photovoltaic Modules for Autonomous Cleaning Robots," Signal & Image Processing: An International Journal, vol. 10, no. 03, pp. 43–59, Jun. 2019, doi: 10.5121/sipij.2019.10305.

[26] T. Cruz-Rojas, J. A. Franco, Q. Hernandez-Escobedo, D. Ruiz-Robles, and J. M. Juarez-Lopez, "A novel comparison of image semantic segmentation techniques for detecting dust in photovoltaic panels using machine learning and deep learning," *Renewable Energy*, p. 119126, Aug. 2023, doi: 10.1016/j.renene.2023.119126.

[27] U. Naeem, K. Chadda, S. Vahaji, J. Ahmad, X. Li, and E. Asadi, "Aerial Imaging-Based Soiling Detection System for Solar Photovoltaic Panel Cleanliness Inspection," Sensors, vol. 25, no. 3, pp. 738–738, Jan. 2025, doi: 10.3390/s25030738.

[28] H. El Karch, R. El Gouri, Y. Natij, M. Benally, and A. Mezouari, "AI-Based Smart Real-Time PV Panels Soiling Recognizing System Using Deep Neural Network Framework on NVIDIA Jetson Nano Embedded GPU," Ingénierie des systèmes d'information, vol. 29, no. 5, pp. 1687–1699, Oct. 2024, doi: 10.18280/isi.290503.

[29] S. Selvi, V. Devaraj, and R. P. P.s, "Detection of Soiling on PV Module using Deep Learning," International Journal of Electrical and Electronics Engineering, vol. Volume 10, Jul. 2023, doi: 10.14445/23488379/IJEEE-V10I7P108.

[30] I. Radonjić, M. A. Amin, M. Petronijević, P. Tsankov, and M. Ćalasan, "Soiling identification and forecasting in urban environment," Engineering Science and Technology, an International Journal, vol. 71, p. 102188, Sep. 2025, doi: 10.1016/j.jestch.2025.102188.

[31] T. Cruz-Rojas, J. A. Franco, Q. Hernandez-Escobedo, D. Ruiz-Robles, and J. M. Juarez-Lopez, "A novel comparison of image semantic segmentation techniques for detecting dust in photovoltaic panels using machine learning and deep learning," Renewable Energy, p. 119126, Aug. 2023, doi: 10.1016/j.renene.2023.119126.

[32] T. Mostafid, "Tumor Segmentation with U-Net and DeepLabv3+: A Review," Medium, Jun. 13, 2024. https://medium.com/@t.mostafid/tumor-segmentation-with-u-net-and-deeplabv3-a-review-048e10001fb2.

[33] A. R. Agduma and R. Dein, "Multi-model convolutional neural network architectures for coastal forest extent and aboveground biomass estimation," Remote Sensing Applications Society and Environment, pp. 101647–101647, Jun. 2025, doi: 10.1016/j.rsase.2025.101647.

[34] M. Kleebauer, C. Marz, C. Reudenbach, and M. Braun, "Multi-Resolution Segmentation of Solar Photovoltaic Systems Using Deep Learning," Remote Sensing, vol. 15, no. 24, p. 5687, Jan. 2023, doi: 10.3390/rs15245687.

[35] X. Chao, L. Zhang, Y. Li, J. Nie, S. Yang, and S. Ercisli, "Optimization Method for Remote Sensing Image-Based Photovoltaic Panel Segmentation via Perception-Driven Enhancement in Nonideal Environments," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 22513-22529, 2025, doi: 10.1109/JSTARS.2025.3602477.

[36] E. Ka, S. Go, Ulziitamir Davaadorj, S. Choi, and A. Nasridinov, "Comparative Analysis of Deep Learning-Based Image Segmentation Models for Solar Panel Soiling Detection," MITA2022, Jul. 2022.

[37] Y. Lei, X. Wang, A. An, and H. Guan, "Deeplab-YOLO: a method for detecting hot-spot defects in infrared image PV panels by combining segmentation and detection," Journal of Real-Time Image Processing, vol. 21, no. 2, Mar. 2024, doi: 10.1007/s11554-024-01415-x.

[38] P. Choure and S. Prajapat, "Exploring U-Net, FCN, SegNet, PSPNet, Mask R-CNN and Using DeepLabV3+ for Multiclass Semantic Segmentation on Satellite Images of Western Ghats," Dec. 2025, doi: 10.22541/au.176479420.03812962/v1.

[39] A. Moghimi, M. Welzel, T. Celik, and T. Schlurmann, "A Comparative Performance Analysis of Popular Deep Learning Models and Segment Anything Model (SAM) for River Water Segmentation in Close-Range Remote Sensing Imagery," in IEEE Access, vol. 12, pp. 52067-52085, 2024, doi: 10.1109/ACCESS.2024.3385425.

[40] D. Sarkar, D. Guha, P. Tarafdar, S. Sarkar, A. Ghosh, and D. Dey, "A comprehensive evaluation of contemporary methods used for automatic sleep staging," Biomedical Signal Processing and Control, vol. 77, p. 103819, Aug. 2022, doi: 10.1016/j.bspc.2022.103819.

[41] M. Carpentier, C. Combescure, L. Merlini, and T. V. Perneger, "Kappa statistic to measure agreement beyond chance in free-response assessments," BMC Medical Research Methodology, vol. 17, no. 1, Apr. 2017, doi: 10.1186/s12874-017-0340-6.

[42] Herlawati and R. T. Handayanto, "Land Cover Segmentation of Multispectral Images Using U-Net and DeeplabV3+ Architecture," Jurnal Ilmu Komputer dan Informasi, vol. 17, no. 1, pp. 89–96, Feb. 2024, doi: 10.21609/jiki.v17i1.1206.

Gas-Efficient Smart Contract Design: Quantifying Refactoring Impact on EVM Execution Costs

Nur Haliza Abdul Wahab[1]*, Juniardi Nur Fadila[2], Nur Faszha Razali[3], Keng Yinn Wong[4]
Department of Computer Science, Faculty of Computing, University Technology Malaysia, Johor, Malaysia [1, 2]
Digital Trust Centre, Nanyang Technological University, Singapore [1]
Department of Informatics Engineering, UIN Maulana Malik Ibrahim, Malang, Indonesia [2]
Kolej Kemahiran Tinggi MARA, Petaling Jaya, Selangor, Malaysia [3]
Faculty of Mechanical Engineering, University Technology Malaysia, Johor, Malaysia [4]

***Abstract*—High transaction costs remain a major barrier to the scalability of Ethereum-based decentralized applications (DApps), particularly when smart contracts are computationally inefficient. Although the Solidity compiler optimizer can reduce bytecode size and improve some low-level patterns, it does not fully address structural inefficiencies in storage layout and state mutation. This study introduces controlled empirical research on the topic of manual smart contract refactoring approaches with the aim of quantifying their impact on gas usage and execution cost in the Ethereum Virtual Machine (EVM). The Remix Integrated Development Environment (IDE) and a synchronized Go-Ethereum (Geth) node (version 1.13.5) were configured to create a controlled experimental environment. This environment was connected to the Sepolia Testnet to approximate conditions similar to the Ethereum Mainnet. The role of high-cost storage operations such as SSTORE was analyzed using opcode-level transaction traces, which were collected using debug_traceTransaction. The proposed refactoring plan implies the alignment of storage slots by systematically packing the variables and data location optimization (calldata and memory) to minimize unnecessary memory allocation. The experiments show gas reductions of up to 40.68% for storage-intensive functions, with an average reduction of 28.5% across all evaluated test cases. Moreover, the findings at the opcode level have shown that it is possible to reduce the costs of unnecessary storage writes without impacting the correct functional performance of the execution. Overall, the findings show that storage-aware manual refactoring is a viable strategy for improving runtime efficiency and reducing the execution cost of Layer-1 smart contracts.**

Keywords*—Ethereum smart contract; solidity refactoring; gas optimization; Ethereum Virtual Machine (EVM); opcode tracing

I. Introduction

Ethereum is a widely used decentralized application platform whose smart contracts are executed through the Ethereum Virtual Machine (EVM) [1], [2]. However, the wider use of Ethereum is still constrained by the gas cost of executing smart contracts [3], [4]. Although gas helps prevent computational abuse, high gas requirements remain a practical barrier to scalability [5], [6]. Ethereum has adopted Proof of Stake (POS) and implemented EIP-1559, which stabilizes market fees. But even the gas that is used by a smart contract is still heavily reliant on the computational and storage requirements of the contract itself. This implementation expense has a direct impact on the cost-effectiveness of Decentralized Applications (DApps) [7], [8].

Despite these protocol-level changes, contract-level code complexity still affects execution cost. This is especially important for high-frequency DApps (such as decentralized order books or real-time gaming), because even marginal inefficiencies in state management will make a project economically unviable for mainstream adoption [9], [10], [11].

While standard Solidity compiler optimizer *'solc --optimize'* as well as automated tools like *'GasReducer'*, *'sOptimize'* make good progress on reducing the size of bytecode using techniques like constant folding and function inlining, they don't explicitly take developer-defined architectural choices, in terms of storage layout and data-handling [12], [13], [14]. Since the compiler has no idea of the organization of storage specific to the contract, it is not able to automatically implement structural optimizations like packing storage slots, such as storage-slot packing and calldata-based parameter handling [15], [16]. This study examines whether manual refactoring can improve runtime gas efficiency in selected Solidity patterns. This motivates attention to storage-aware refactoring strategies.

To address such an efficiency difference, a rigid comparative empirical analysis is made in this study between the traditional automated methods of optimization and advanced manual refactoring techniques. The method applied in this study to examine the cost of execution in detail is the opcode-level transaction tracing. This is especially so in applications that are highly interactive in the blockchain. Based on this, the aims of the study are the following:

- Implement a manual refactoring framework for Solidity contracts, focusing on storage slot alignment (Variable Packing) and data location calldata vs. Memory.
- Conduct comparative empirical research to quantify the gas savings when deployed and run-time execution with respect to the case of the baseline and optimized versions.
- Determine the deterministic behavior of EVM through trace fine-grained opcodes with a synchronized private Geth node.

To further direct the empirical study, the following research questions are derived in this study:

*Corresponding author.
This research was supported by a UTM Fundamental Research Grant Q.J130000.3828.23H38).

- RQ1: Does storage-slot alignment and data-location optimization of manual refactoring decrease gas consumption of storage-intensive Solidity functions when controlled by EVM execution?
- RQ2: Which optimization strategy contributes more directly to gas reduction at the opcode level: variable packing or calldata-based parameter handling?
- RQ3: How significant are the observed gas savings in terms of meaningful economic benefits to high-frequency cases of decentralized application?

These research questions are intended to provide a more analytical framework for the structure of the assessment of the technical and economic impact of manual smart contract optimization.

This study empirically evaluates the manual refactoring workflow for reducing EVM execution cost under controlled test conditions. By presenting at the opcode level evidence of runtime gas savings up to 40.68%, this study shows that manual interventions in storage-slot alignment and data-location management can improve execution efficiency and lower transaction-level cost for the evaluated smart contract functions:

- An opcode-level empirical framework for analyzing gas optimization in Ethereum smart contracts within a controlled EVM experimental environment.
- A manual Solidity refactoring methodology targeting storage access patterns and calldata optimization to reduce execution costs.
- Experimental validation demonstrating gas reductions of up to 40.68% through structural code refactoring under controlled Sepolia Testnet conditions.

II. Literature Review

Previous studies on the efficiency of Ethereum smart contracts can be classified into four categories: gas-fee dynamic, EVM resource pricing, automatic optimization tools, and experimental validation environments [17], [18], [19]. However, there is little evidence in the literature on the impact of manual refactoring on execution cost in controlled EVM environments at the level of opcodes [11], [18]. This gap indicates the need for regulated empirical assessment of manual refactoring utilizing low-level execution traces.

A. Economic Scalability and Gas Volatility

Recent protocol changes have not removed the importance of contract-level gas efficiency [12], [20]. Some studies say that the cost of gas is a big problem for DApps since bugs in contract code make transactions more expensive for users [21], [22].

Under EIP-1559, transaction fees are divided into a protocol-determined base fee and a priority fee [23]. Although fee levels vary with network conditions, developers still influence overall transaction cost through the gas consumption of contract execution [16], [17]. Because fee levels are externally determined, developer influence lies primarily in reducing contract-level gas consumption.

To frame optimization more clearly, gas mechanisms can be categorized according to their functions within the blockchain system [24]. In Ethereum, gas serves two main purposes: it acts as a pricing mechanism that discourages computational abuse, and it helps mitigate unbounded execution and denial-of-service risks [9], [23]. The given mechanism can be further divided into a two-tiered taxonomy of source control:

1) Protocol control (external): This includes the Base Fee calculated by the EIP-1559 protocol depending on block congestion [6], [23]. Although L2 networks like Polygon or Optimistic Rollups solutions have reduced fee taxonomies due to compression of the data, the logic behind the execution is still governed by the EVM rules [25].

2) Developer source control (internal): This is the strategic control of storage and use of computational resources at the implementation level [3], [4]. Although fee-market charges under EIP-1559 include a Base Fee and a Priority Fee, developers primarily influence total gas usage through contract-level efficiency, as illustrated in Fig. 1. Even though developers are not in control of the Base Fee in the market, they have ultimate control over gas limit through the efficiency of their code [21].

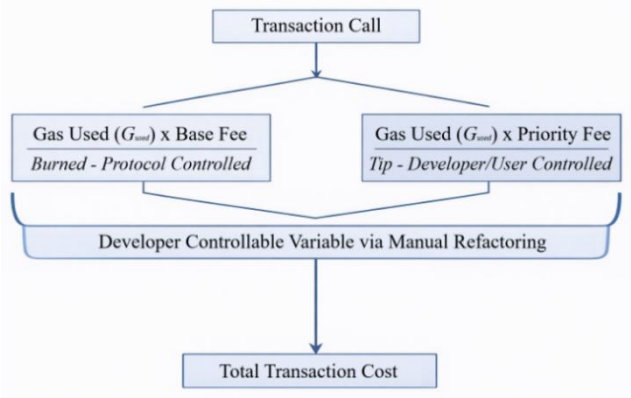


Fig. 1. Conceptual framework of Ethereum gas fee calculation.

Prior studies show that private Geth nodes support controlled and reproducible measurement of EVM execution behavior, making them suitable for fine-grained gas analysis [5], [26].

B. EVM Operating Cost Calculation Mechanism

The deterministic aspect of EVM calculation of costs is a requirement before one can detect structural inefficiencies in smart contracts [10], [18]. Reducing code inefficiency remains important for improving execution efficiency. Each operation is metered to support network security and to prevent unbounded execution, including risks associated with infinite looping [10], [27]. The total gas used by a transaction G_{total} is a combination of computational, storage, and memory resources usage, that is controlled by Eq. (1):

$$G_{total} = G_{intrinsic} + G_{execution} + G_{memory} \quad (1)$$

TABLE I. GAS-INTENSIVE EVM OPCODES AND OPTIMIZATION TARGETS

Opcode	Typical Gas Cost	Resource Category	Optimization Relevance	Refactoring Target
SSTORE	20,000 (New) / 2, 900 (Dirty)	Persistent storage writes	Very High	Variable Packing
SLOAD	2, 100 (Cold) / 100 (Warm)	Persistent storage read	High	Reduce repeated reads
MLOAD / MSTORE	3	Memory	Low	Temporary caching
CALLDATACOPY	3 + 3 x words	Input data copy	Moderate	Calldata vs. Memory
ADD / SUB	3	Stack Arithmetic	Low	Basic Logic

where:

- $G_{instrinsic}$: The fixed 21,000 units of gas to pay the base cost per transaction contained in the block.
- $G_{execution}$: The total cost of the opcodes that have been performed at runtime. This consists of state operations of high cost, such as *SSTORE* (up to 20,000 gas) and of low cost, such as *ADD* (3 gas).
- G_{memory}: Costs that are paid in quadratic proportion to the increase in memory used in contract execution.

Analysis through Eq. (1) reveals that the most impactful optimization vectors lie within $G_{execution}$ and G_{memory}. For instance, [20] highlight that identifying costly bytecodes and replacing them with gas-efficient patterns is highly required to mitigate these expenses. Techniques such as Variable Packing directly target the reduction of $G_{execution}$ by minimizing the expensive of storage operations, while shifting data location to *CALLDATA* mitigates the G_{memory} expansion penalty.

However, as mentioned by [5], the issue of linking these technical assurances to particular audit statements is the focus of research. Table I has been synthesized as comparative technical specifications of these opcodes, which are resource-intensive.

Table I summarizes the gas-expensive EVM opcodes that are most relevant for this study. In particular, *SSTORE* dominates state-changing cost, making storage-slot packing a key target for reducing deployment and runtime overhead. By contrast, input-copy operations such as *CALLDATACOPY* highlight the importance of parameter handling through calldata to avoid unnecessary memory-expansion costs. Collectively, these characteristics of the opcodes give the technical underpinning for the refactoring strategies that are evaluated in this work.

On top of this classification, Section III introduces the experimental refactoring workflow that is used to baseline smart contract implementations.

C. Manual Refactoring vs Automated Tools

Compiler-level optimization can improve common low-level patterns, but it does not usually perform structural changes such as storage-slot reorganization or context-specific data-layout decisions [28].

In order to overcome these shortcomings, different specific automated tools have been produced. *GasReducer* was presented in [15] as a tool that recognizes 24 code anti-patterns using the bytecode representation, and substitutes them with more efficient operation sequences.

TABLE II. COMPARISON OF SMART CONTRACT GAS OPTIMIZATION APPROACHES AND THEIR LIMITATIONS

Optimization Approach / Study	Primary Focus	Core Technique / Mechanism	Strengths / Contributions	Critical Limitation (Efficiency Gap)
Solidity Compiler Optimizer (--optimize) [29].	Bytecode size reduction, basic runtime optimization	Deduplication, jumpdest removal, peephole optimizations	Provides baseline automated improvements with minimal developer effort	Cannot restructure storage layout or apply variable packing; limited awareness of opcode-level cost (e.g., *SSTORE*)
GasReducer [30]	Anti-pattern detection at bytecode level	Replaces inefficient opcode sequences (24 patterns)	Effective for removing known inefficiencies automatically	Pattern-based; lacks semantic context and cannot perform deep architectural refactoring
sOptimize [31].	Dead code and redundancy elimination	Control Flow Graph (CFG) + lazy annotation	Improves execution efficiency through redundant instruction removal	Limited ability to optimize storage-heavy functions or calldata transitions
GASOL [2], [18], [32].	Minimizing memory/storage operations	SMT-based optimization encoding	Formal approach to reduce expensive memory and storage usage	Computationally complex; does not provide developer-guided storage slot packing
Slither Refactoring Framework [33]	Static detection of storage inefficiencies	Static analysis of contract structure	Identifies refactoring opportunities and vulnerability patterns	Context-blind: cannot apply dynamic optimizations such as calldata vs. memory restructuring
GasOptiScan (AST-based tools) [34].	Loop and syntax-level optimization	Abstract Syntax Tree (AST) parsing for pattern matching	Useful for detecting loop-fusible or structural patterns	Cannot optimize calling-context-dependent transitions (e.g., memory → *CALLDATA*)
Audit Formalization Gap [35].	Verification and guarantees mapping	Linking optimization guarantees to audit assertions	Highlights need for stronger runtime-focused validation	Automated tools prioritize deployment reduction over runtime opcode efficiency
Manual Refactoring Framework Evaluated in This Study	Runtime execution efficiency at opcode level	Variable packing, calldata optimization, stack caching, opcode tracing via debug traceTransaction	Provides opcode-level empirical evidence of runtime gas savings through storage-aware restructuring	Requires deeper EVM expertise and manual developer intervention (higher barrier to entry)

TABLE III. COMPARATIVE ANALYSIS OF EXPERIMENTAL ENVIRONMENTS FOR GAS VALIDATION

Blockchain Network	Consensus Mechanism	Avg. Transaction Cost (USD)	Throughput (TPS)	Cost Implication for Developers
Ethereum (Layer 1)	Proof-of-Stake (PoS)	$2.50 - $15.00	~15 -30	Severe: Code optimization is mandatory for viability
Arbitrum One (L2)	Optimistic Rollup	$0.10 - $0.50	~40,000+	Moderator: Cheaper, but inherits L1 data availability costs
Solana	Proof-of-History (PoH)	< $0.001	~65,000	Negligible: Optimization is less critical due to abundance.
Sepolia Testnet	Proof-of-Stake (PoS)	Free (Testnet ETH)	Mirrors Mainnet	Suitable: Useful for controlled testing of L1 logic without direct financial risk

On the same note, [30] created *sOptimize*, a tool that uses Control Flow Graph (CFG) representations and lazy annotation to identify dead or redundant code [4]. On a further level, [2], [18], [32] suggested GASOL, a tool that uses Satisfiability Modulo Theories (SMT) encoding to reduce memory and storage operations. Moreover, [33], [36], used Slither framework to find refactoring scenarios in the storage usage, showing that automated static analysis can conserve an average of 41,184 gas per contract.

Although these tools have been useful, there is a long-standing efficiency gap as a result of the "context blindness" inherent to them [4]. Although analysis tools such as *GasOptiScan* make use of Abstract Syntax Tree (AST) parsing to determine loop-fusible patterns, which fail to achieve the complexity structural restructuring (such as the strategic reorganization of memory into *CALLDATA* locations) in large arrays or structs that are not updated, it is shown that rearranging the parameters of a function to *CALLDATA* can yield substantial gas savings [15].

Few studies provide opcode-level empirical validation of manual refactoring strategies under controlled EVM conditions. Automated tools tend to optimize deployment size rather than the cost of runtime execution. However, few studies provide opcode-level empirical validation of manual refactoring strategies within controlled EVM environments, leaving a gap in understanding how structural code changes directly affect gas consumption during execution.

This gap motivates the present study, which evaluates manual refactoring using opcode-level tracing to examine how storage-aware changes affect runtime gas consumption. Table II indicates that automated tools are useful for common inefficiencies but remain limited for storage-aware and context-dependent runtime optimization. This supports the need for controlled opcode-level evaluation of manual refactoring.

D. From Economic Volatility to Deterministic Validation

Public-network measurements may be affected by latency, congestion, and other sources of noise, which can obscure fine-grained analysis of execution cost [37]. Recent studies increasingly favor controlled experimental environments for reproducible gas evaluation (Table III) [13], [38]. However, tool-based reports alone do not fully capture how manual refactoring changes internal execution behavior [9], [13]. Sepolia is increasingly used as an Ethereum-compatible test environment for gas evaluation because it preserves the EVM execution model while avoiding real-market cost variability [39].

When combined with private Geth tracing, Sepolia-compatible evaluation provides a reproducible basis for opcode-level analysis of execution behavior [38], [39], [40].

III. METHODOLOGY

This section outlines the experimental design used to evaluate the effectiveness of the manual Solidity code-optimization techniques. Rather than relying solely on abstract simulation, the study uses a deterministic experimental setup based on synchronized local Geth nodes. This is important to ensure that the dependent variables that are being isolated are gas consumption $G_{totalused}$ and opcode execution costs, and that these are independent of network noise (randomness of public latency or mempool front-running (MEV) interference).

The experimental protocol was organized in a workflow to help with data integrity, reproducibility, and procedural rigor, as shown in Fig. 2. The methodology was put in place in four connected stages:

1) Phase I-Environment setup: Initializing Geth v1.13.5 nodes with custom genesis parameters to approximate Ethereum Mainnet-like execution conditions.

2) Phase II-Contract development: Building two smart contract variants (baseline and optimized) using manual refactoring techniques.

3) Phase III-Transaction execution and mining: Execute standardized workloads through Remix IDE connected to a private network. This phase involves the determination of function execution using three experimental controls. To ensure consistency at the baseline, the compiler version of Solidity v0.8.20 (Paris hardfork) and the compiler optimizer (runs: 0) were used for both Cbase and Copt to isolate the direct effect of manual refactoring and automated compiler improvement. To improve measurement stability, each function was run for 50 runs to ensure the data is consistent and the "cold-access" outliers are removed, which could skew the average amount of gas consumed. Finally, the issue of input variation was solved by standardized workloads, and for calldata optimization, the dynamic payloads of 32, 64, and 128 bytes were tested to investigate the performance of various input sizes.

4) Phase IV-Data extraction: Use of debug_traceTransaction API call to obtain granular data at the opcode level for comparative analysis.

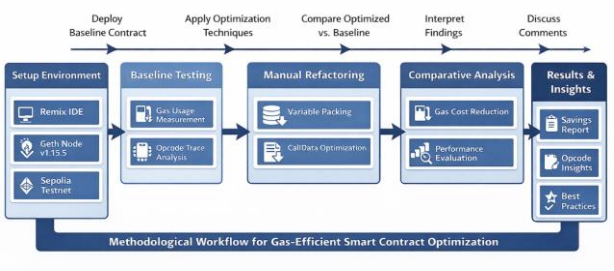


Fig. 2. Methodological workflow for smart contract optimization.

While Fig. 2 is an overview of the methodological stages, the experimental execution architecture is presented in Fig. 3. This figure shows the data flow between the user interface, wallet layer, RPC interface, and the underlying Geth execution client in the private network.

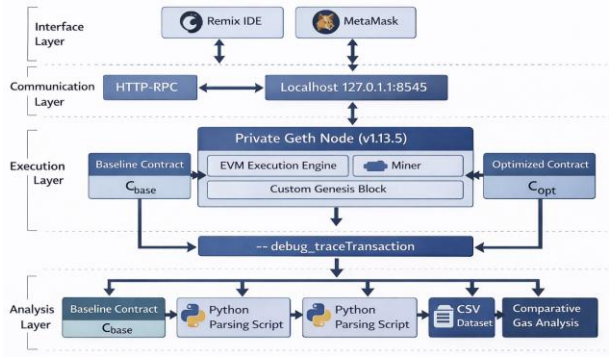


Fig. 3. Architecture of the experimental environment.

This architecture supports closed-loop opcode tracing while minimizing public-network noise.

A. Experimental Architecture and Environment Configuration

The experiments were carried out in a hybrid development environment that had the Windows interface and a Linux execution layer in order to support steady measurement of gas-related execution behavior. Such architecture had been designed to rule out such confounding factors as the network lag and the fluctuations of the global gas market.

To aid the reproducibility of the results, the experimental setup was run on a workstation with the AMD Ryzen 7 5800H CPU, 16 GB DDR4 RAM, and 512 GB NVMe SSD storage. The software stack used Windows Subsystem for Linux 2 (WSL2) that runs Ubuntu 22.04 LTS. A private node on Ethereum was set up using Go-Ethereum (Geth v1.13.5). The node was set up with a custom genesis block, which set the gasLimit to 30,000,000 units. This configuration created an isolated private network as shown in Fig. 4.

```
nurhaliza@DTC-D3Q0R14:~/Solidity$ geth version
Geth
Version: 1.13.5-stable
Git Commit: 916d6a441a866cb618ae826c220866de118899f7
Git Commit Date: 20231114
Architecture: amd64
Go Version: go1.21.4
Operating System: linux
GOPATH=
GOROOT=
```

Fig. 4. Geth v1.13.5 used for the experimental environment.

To isolate logic execution cost from fee effects, the miner was configured with a zero-gas-price policy (*--miner.gasprice 0),* allowing transactions to be processed without market-driven gas-price variation. The next set of command line instructions are executed in the Linux setting as indicated in Fig. 5.

```
nurhaliza@DTC-D3Q0R14:~/Solidity$ geth --datadir ./node1 --http --http.api "eth,net,web3,pe
rsonal,debug" --http.corsdomain "*" --networkid 12345 --allow-insecure-unlock --miner.gaspr
ice 0
ctory"
INFO [02-12|13:20:19.249] Set global gas cap                       cap=50,000,000
INFO [02-12|13:20:19.250] Initializing the KZG library             backend=gokzg
WARN [02-12|13:20:19.263] Sanitizing invalid miner gas price       provided=0 updated=1,000,000,000
INFO [02-12|13:20:19.264] Allocated trie memory caches             clean=154.00MiB dirty=256.00MiB
INFO [02-12|13:20:19.264] Defaulting to pebble as the backing database
INFO [02-12|13:20:19.264] Allocated cache and file handles         database=/home/nurhaliza/Solidity/node1/geth/chaindat
a cache=512.00MiB handles=524,288
INFO [02-12|13:20:19.285] Opened ancient database                  database=/home/nurhaliza/Solidity/node1/geth/chaindat
a/ancient/chain readonly=false
INFO [02-12|13:20:19.285] State schema set to default              scheme=hash
INFO [02-12|13:20:19.285] Initialising Ethereum protocol           network=12345 dbversion=<nil>
INFO [02-12|13:20:19.286] Writing default main-net genesis block
INFO [02-12|13:20:19.475] Persisted trie from memory database      nodes=12356 size=1.79MiB time=52.961898ms gcnodes=0 g
csize=0.00B gctime=0s livenodes=0 livesize=0.00B
INFO [02-12|13:20:19.501]
INFO [02-12|13:20:19.501] ---------------------------------------------------------------------------------------------
---------------------------------------------------------
INFO [02-12|13:20:19.501] Chain ID:  1 (mainnet)
INFO [02-12|13:20:19.501] Consensus: Beacon (proof-of-stake), merged from Ethash (proof-of-work)
INFO [02-12|13:20:19.501]
INFO [02-12|13:20:19.501] Pre-Merge hard forks (block based):
```

Fig. 5. Command line for the node execution.

The following technical reason was used to pick this flag formation:

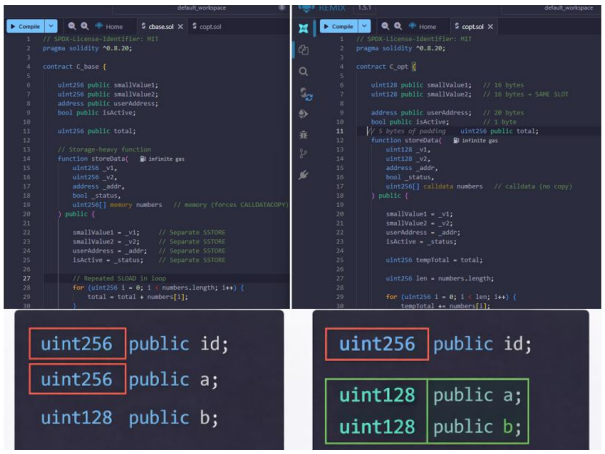
- *--http.api"...,debug"*: Enables the *debug*_traceTransaction module, which is required to extract granular opcode-trace data (structLogs) for the analysis.
- --miner.gasprice, 0: removes market-driven price variability, allowing the analysis to focus on gas used, Gused (code efficiency), rather than gas price, Pgas (market conditions).
- Communication between Remix IDE and Geth node was done via 127.0.0.1, which reduces the communication delay locally and provides more consistent execution-time observation.

B. Contract Development and Refactoring Design

To examine the design decision impact on the cost of execution, the present study relies on a paired-comparison model. Two variants of smart contracts were created with similar functional behavior, but different internal structural designs:

- Baseline Contract, C_{base}: Represents standard coding practices without any gas-saving strategies, often found in legacy or new developer code.
- Optimized Contract, C_{opt}: Functionally equivalent code but has been through an intensive manual refactoring process using gas-oriented design techniques.

As visualised in Fig. 6, the progression from C_{base} to C_{opt} follows the principle of semantic equivalence. Both contracts should give the same observable output for the same input x and the same intended state transition, $f(C_{base}, x) = f(C_{opt}, x)$. This condition makes sure that any reduction of gas is due to refactoring of the architecture and not due to a change or reduction in functionality of the contract.



(a) (b)

Fig. 6. Comparison of structure between (a) Baseline contract vs (b) Optimized contract.

This refactoring model will consider two essential optimization mechanisms that address definite inefficiencies in the EVM:

1) Storage slot packing (variable alignment): The EVM operates using a 256-bit (32-byte) word size. In the C_{base}, variables were arranged naively, for example, by using unnecessarily large types or suboptimal ordering so that storage slots were not efficiently utilized. This causes repeated opcodes of SSTORE, which is the most expensive operation (20000 gas to write cold). The Variable Packing technique is used in the C_{opt} contract. Small variables (including uint128, address, and Boolean) are ordered contiguously so that they can be placed in a 32-byte slot. Mathematically, in case S is the number of storage slots in use, as in Eq. (2).

$$S_{opt} < S_{base} \quad (2)$$

Under naive allocation, S_{base} requires a larger number of slots, whereas S_{opt} reduces slot usage through efficient ordering and packing of compatible variables.

2) Data location optimization: Memory vs Calldata: For functions that accept reference-type parameters (such as arrays or strings), Cbase uses the memory keyword. This forces the EVM to copy the argument data from calldata into temporary memory (memory expansion), which incurs additional gas costs for memory allocation and opcode copying. In contrast, C_{opt} uses the calldata keyword where applicable. By taking the location of the data to be the location of the function, which is given in the value of the CALLDATA, the data is read directly from the input of the transaction, without being copied to memory. This not only reduces data-copy overhead and the memory-expansion cost, but also avoids the quadratic memory expansion cost C_{mem} that is calculated:

$$C_{mem} = 3a + \left(\frac{a^2}{512}\right) \quad (3)$$

where, a is the size of the allocated word of memory.

C. Data Acquisition and Opcode-Level Tracing

In order to reach accurate determinations of the cost of execution over and above the aggregate gas numbers available through the high-level interfaces, a direct Remote Procedure Call (RPC) connection was established between the Remix IDE and the local Geth node. Instead of using the injected *window.ethereum* provider (Metamask) that causes the network latency and gas estimation buffers, the Remix IDE was set up to communicate directly with the local Geth node using the HTTP-RPC protocol on port 8545. This configuration enables the direct injection of signed transactions into the private mempool, ensuring that the execution environment remains strictly deterministic.

On successful mining of each transaction (for both C_{base} and C_{opt}, the transaction has (TXhash) captured. To get granular data of execution, in this research, Geth's built-in debug API has been used. The API returns a raw structure of the *structLog*, a chronological sequence of all the opcodes that are executed by EVM in the form of a RAW JSON Object.

As shown in Fig. 7, this log gives important metrics about each line, including the Program Counter (PC), Opcode Name, Gas Cost, and Stack Depth. Data from these logs was parsed to specifically filter the *SSTORE* (storage write) and *SLOAD* (storage read) operations in order to be able to perform a comparative cost isolation between the baseline and optimized contracts.

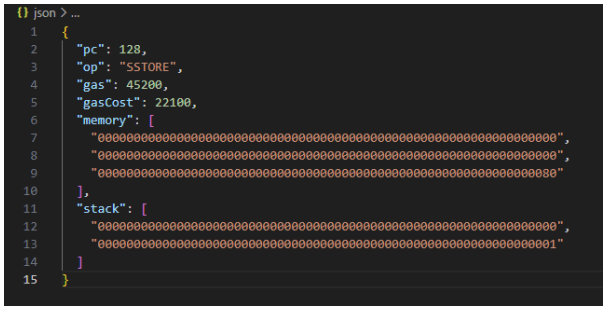


Fig. 7. Command line invocation for trace extraction.

To make the analyses analytical with accuracy, raw datasets of the JSONs were processed using a Python script that was written specifically for this analysis. This automated pipeline removed non-target overhead by separating the contract's business-logic execution cost from the intrinsic transaction fee (21,000 gas) and converted the unstructured traces into structured CSV files for comparative analysis.

To check the stability of the measurements, each of the tested functions was performed multiple times under the same conditions. Since the observed gas values were unchanged from run to run, the analysis was directed towards direct comparison of gas and reproducibility as opposed to inferential statistical testing.

D. Performance Metrics and Evaluation Model

To tackle the technical trace data and to achieve meaningful performance indicators, in this study, a quantitative evaluation model was adopted. This model measured the effectiveness of optimization using three main metrics: absolute gas

consumption, financial cost implications, and relative efficiency percentage.

The basic metric used is the number of units of gas used in the successful execution of transactions. This value is taken directly from the G_{used} field in the receipt of a transaction. It represents the actual amount of computation that the EVM will work on without any external market factors. Since the main objective of the study is the economic scalability, the gas consumption G_{used} is converted to fiat monetary value (USD) to reflect the actual cost that the end user has to bear. Transaction cost was determined from Eq. (4):

$$C_{tx} = (G_{used} x\ P_{gas} x\ 10^{-9}) + P_{eth} \quad (4)$$

where, P_{gas} is the average gas price in Gwei units, while P_{eth} is the current Ether market price in USD. The values of P_{gas} and P_{eth} are adjusted based on the market average during the period of data collection in order to simulate the realistic scenarios of the Mainnet.

To quantify the scale of savings done through techniques of manual refactoring, the Optimization Efficiency Ratio metric was introduced in this study. This metric compares the gas performance of the optimized contract $G_{optimized}$ against the baseline contract $G_{baseline}$. The equation for the calculation of the savings percentage Δ_{gas} is written in Eq. (5):

$$\Delta_{gas} = \left(\frac{G_{baseline} - G_{optimized}}{G_{baseline}}\right) x\ 100\% \quad (5)$$

The positive value of the Δ_{gas} is a sign of reduced computational burden; a value close to zero is a sign of the ineffectiveness of a particular optimization technique. This metric can be used to make direct comparisons of the effects of variable packing and calldata optimization for different contract functions. This evaluation framework also ensures that the findings are not just technically based but that they are also economically relevant to DApps development.

A. Methodological Flowchart of the Study

To ensure the clarity of the procedures and reproducibility of the study, the entire design of the study is summarized in a series of steps as shown in Fig. 8.

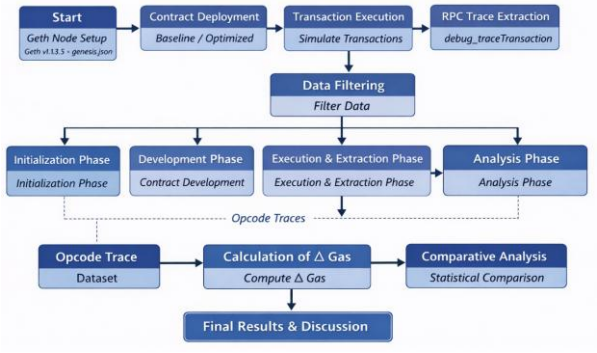


Fig. 8. Comprehensive methodological workflow of the study.

This flowchart shows the procedure in the study from the environment configuration in nodes to the final analysis and synthesis. The process has been broken down into 4 main interrelated phases:

- Initialization Phase: Here, the preparation of a Geth v1.13.5 node with a custom genesis.json file will be done to provide a deterministic EVM environment.
- Development Phase: Two versions of the smart contracts (baseline vs. optimized) will be coded in the Remix IDE and the Stonecode compiled.
- Execution & Extraction Phase: Execution of simulated transactions and extraction of raw opcode data via the *debug_traceTransaction* API.
- Analysis Phase: A trace data filtering, performance measures analysis, and the gas-saving effects analysis in comparison.
- The data flow is indicated between the software components as well as the flow of the data to represent the end-to-end integrity of the experimental procedure.

B. Threats to Validity

Although this study was meant to provide a controlled and reproducible evaluation of the optimization of manual smart contracts, there are some validity considerations to be acknowledged.

Internal validity was taken care of by running both the baseline and optimized contracts under an identical experimental environment, compiler configuration, and measurement procedure. Repeated executions were performed to minimize random variation in the gas observation. However, some uncontrolled factors at the implementation level might have affected the measured gas costs, e.g., function-specific execution paths, storage access patterns, and EVM runtime behaviour.

Construct validity, which is concerned with whether the chosen metrics are appropriate to selected optimization benefits under investigation. In this work, gas consumption and op-code level traces were used as the main indicators for the computational efficiency. These measures are suitable for determining the cost of execution in Ethereum smart contracts, but they are not sufficient to determine other software quality dimensions such as readability, maintainability, auditability, or complexity of long-term upgrades.

External validity is limited by the range of the evaluated contracts and execution situations. The study focuses on a controlled set of Solidity functions and optimization patterns, particularly storage-slot packing and calldata-based parameter handling. Therefore, the results cannot necessarily be generalized to all the different kinds of smart contract architectures, all the different blockchain platforms, and production environments with more complex transaction flows, more multi-contract interactions, or evolving gas rules.

Despite these limitations, the controlled experimental design provides useful evidence on the gas-saving potential of the manual refactoring strategies in Solidity-based smart contracts. Following this methodology framework and associated validity considerations, the results of the experiment are given in the next section.

IV. Result and Discussion

This section presents the empirical results that are obtained from the transaction receipts and opcode traces of the controlled environment Geth. The trace data were filtered in order to remove the interrupting effect of the non-overhead-related cost of contract execution and to minimize the interference of external network conditions. The results are discussed in three analytical dimensions, namely, quantitative performance, behavior at the level of opcodes, and economic implications:

Quantitative Performance Analysis: Quantitative analysis of the magnitude of the reduction of gas consumption by comparing the baseline and optimized contracts at the level of the total gas savings Δ_{gas}. The main concern is with the measurement of the gas unit reduction in the critical functions.

Granular Opcode Interpretation: Describes what made the seen savings possible with the help of execution traces, focuses on reducing storage write operations, and modifies the behavior related to memory.

Economic Impact Assessment: Converts gas savings into estimated cost savings under selected gas-price scenarios to illustrate the practical financial implications of the proposed refactoring.

The section ends with a discussion of the trade-off between efficiency and maintainability, plus a short interpretation of the obtained findings in relation to select Ethereum protocol incentives.

A. Quantitative Performance Analysis

Comparative Analysis between the baseline and the optimized contracts revealed the reduced gas consumption of the optimized version of the contract for all the tested functions under the controlled EVM environment. The results show that variable packing and calldata optimization mitigate the amount of gas consumed and modify the profile of the opcodes and storage accesses of the evaluated contract functions.

For the storage-intensive function *storeData(),* the baseline contract used on average 105 200 gas units. Following the storage-slot packing method, the optimized contract needed 62,400 gas units to perform the same logic. Using the efficiency ratio of Eq. (5), the gas saving for *storeData()* is 40.68%:

$$\Delta_{gas} = \left(\frac{105{,}200 - 62{,}400}{105{,}200}\right) x\ 100\% \approx 40.68\%$$

The larger saving for *storeData()* is due to the nature of the packing of storage slots. For compatible variables in the baseline contract, they are stored in different storage slots, which makes them more expensive to write to storage during execution. An optimized contract requires less storage-write overhead, as compatible variables are packed into the same slot, which results in a 40.68% reduction in gas. On the other hand, *updateRecord()* takes advantage of memory to calldata parameter changing. This optimization primarily eliminates copying and memory-expansion overhead, without eliminating the same amount of storage-write cost. Thus, it is expected that the lower reduction of 15.40% is achieved due to optimization of calldata, while the packing of storage slots is limited to more expensive persistent storage operations. The gas consumption for each function is shown in Table IV for both the baseline and optimized contracts.

TABLE IV. Function-Level Gas Consumption Comparison Between Baseline and Optimized Contracts

Function	Baseline Gas (Units)	Optimized Gas (Units)	Δ_{gas} Savings (%)
storeData()	105,200	62,400	40.68%
updateRecord()	52,000	43,992	15.40%
initializeState()	85,000	61,200	28.00%
processPayment()	48,000	33,600	30.00%
Average Saving	72,550	50,298	28.50%

Note: Values were the same for different runs with the same compiler settings, inputs, and local node conditions.

Fig. 9 is a visual comparison of the gas consumption in the tested functions, and there is a consistent efficiency gap between the baseline and optimized contract variants.

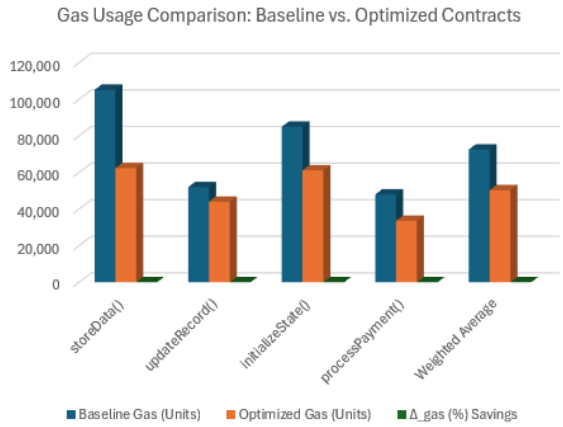


Fig. 9. Gas-use comparison (baseline vs optimized) across key functions.

The average gas savings of all of the functions considered is 28.50%. Conditions tested in Fig. 9 have shown that the optimized contract was always superior to the baseline contract.

To measure the stability of the measurement, each function tested was executed repeatedly with the same compiler options, optimizer level, contract code, contract state, input payload, and local Geth execution environment. With these values fixed, the gas used in execution was the same for repeated runs, and so were the mean, min, max, and standard deviation values, with a standard deviation of zero. This is not to imply fixed network transaction costs. Instead, it's a sign that the code executed by the same function call was still deterministic in the controlled experimental setup! The reported values are thus based on execution gas from the configured Geth tracing environment, rather than pre-computed values or values based on the market prices of transactions.

B. Granular Opcode Interpretation

Although the above section had measured the net drop in gas consumption, strict validation of the suggested refactoring framework cannot be done without the analysis on the execution-trace level. Transaction traces were obtained to analyze the structural origin of such savings by the

debug_traceTransaction module of a synchronized node of a private Geth at version 1.13.5. This allowed direct observation of opcode frequency and gas contribution within the EVM execution state.

The results of the trace show that the savings that have been observed cannot be explained by the decrease in code length or simplification of the control flow. In place, the optimized contract uses a special strategy of substitution where certain expensive storage operations are substituted with cheap stack-based and bitwise instructions. This transformation modifies the opcode composition of the execution trace while preserving functional equivalence. Table V presents the trace-derived opcode frequencies and their associated gas contributions for both the Baseline and Optimized contracts.

TABLE V. TRACE-DERIVED OPCODE FREQUENCY AND GAS CONTRIBUTION

Opcode	Frequency (Base)	Frequency (Opt)	Gas Cost (Base)	Gas Cost (Opt)	Δ_{gas} Difference (Units)
SSTORE	5	2	100,000	40,000	-60,000
SLOAD	4	2	8,400	4,200	-4,200
CALLDATACOPY	0	1	0	6	6
SHL/SHR	0	8	0	24	24
OR/AND	2	6	6	18	12
Other (Stack)	85	92	255	276	21
Total Trace	108	115	108,697	44,536	64,161

Note: Gas totals in this table refer to traced opcode contributions and may not exactly match full transaction gas reported in Table IV.

As Table V demonstrates, the implemented baseline had five SSTORE operations to store independent variables in different storage slots of 256 bits, which cost around 100,000 gases in storage. This was optimized by the optimized version to two SSTORE calls when used with variable packing, where the multiple logical values are packed into aligned storage slots.

In order to support this packing mechanism, the optimized contract added 12 bit-wise operations (SHL and OR). The aggregate cost of these extra stack-level instructions is insignificant, however (12 x 3 = 36 gas) compared to the cost of 20,000 gas of a single SSTORE. This would then give a net write of three storage eliminations, translating to about 60,000 gas loss, even though there would be a small change in the number of instructions.

This evidence of the trace level indicates that the efficiency of smart contracts is controlled mostly by the distribution of opcodes and not the total number of executed instructions. Storage-bound operations carry an over-represented high economic cost compared to CPU-bound stack manipulations according to the gas counting model of Ethereum. The proposed refactoring strategy enables a reduction in storage intensity but does not affect semantic correctness by strategically replacing the state mutations with deterministic bitwise transformations.

Fig. 10 visualizes this structural shift by comparing opcode gas contributions between the two implementations. The pre-emption of *SSTORE* in the profile of baseline execution is greatly minimized in the optimized contract, and the extra penalty used by the bitwise logic is not much compared to it.

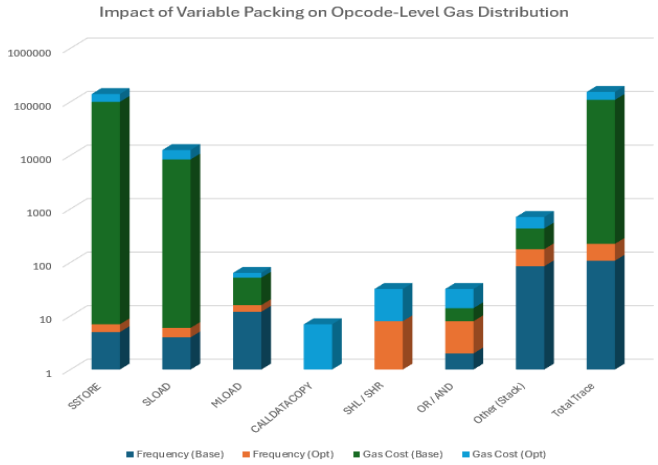


Fig. 10. Opcode gas contribution in baseline and optimized contract execution.

In general, the trace analysis indicates that the savings in gas that is observed are the result of architectural restructuring but not of superficial syntactic adjustment. The opcode traces are a direct indication that the refactoring modifies runtime behavior predictably and understandably.

In order to provide a contextualization of these opcode-level results, Table VI analyses a comparison of the current study to the other empirical studies concerning the optimization of smart contract gas. The comparison reveals the differences in the areas of optimization focus, validation basis, and depth of analysis and demonstrates that the current work not only offers quantifiable savings in gas but also provides deterministic opcode-trace evidence on how such savings are made.

Previous studies, as indicated in Table VI, have demonstrated the utility of the re-writing of the code in bytecode, the optimization through the use of frameworks for static analysis, and automated optimization. Nevertheless, most of these strategies focus on detection or code-level transformation with no controlled runtime attestation of the behavioral change of storage-related opcodes in response to refactoring. In contrast, the present study links function-level gas reduction directly to opcode-trace behavior in a private Geth environment, thereby strengthening the interpretability of the observed optimization gains.

C. Economic Impact Assessment

Because the price of gas fluctuates over time, as does the price of ETH, the actual observed gas savings also have scenario-dependent economic implications. This study uses three network congestion scenarios to analyze the financial impact of network congestion: Low (15 Gwei), Average (30 Gwei), and High (100 Gwei), assuming the base price of Ether is set at $2,500 USD. Based on experimental data, the difference in the gas used by the Baseline and Optimized contracts is on the order of 42,800 gas units per *storeData()* transaction.

TABLE VI. COMPARISON OF THE PRESENT STUDY WITH RELATED WORK ON SMART CONTRACT GAS OPTIMIZATION

Study	Focus	Validation Basis	Reported Outcome	Main Gap Addressed by This Study
GasReducer [15]	Bytecode anti-pattern	Deployed contract and trace analysis	Detects inefficient opcode patterns automatically	Does not capture developer-guided storage restructuring or calldata-aware refactoring
sOptimize [14]	Redundant instruction elimination	CFG-based static optimization	Improves execution efficiency through code simplification	Limited support for storage-heavy runtime restructuring
GASOL [38]	Memory/store age cost reduction	Formal optimization encoding	Reduces selected EVM resource costs	Formal but less oriented toward explainable manual source-level refactoring
Slither-based refactoring [41]	Static storage inefficiency detection	Static structural analysis	Identify refactoring opportunities	Does not validate runtime savings through opcode-trace evidence
AST-based tools [42]	Loop and syntax-level optimization	Source-code pattern analysis	Detects structural inefficiencies automatically	Limited for context-dependent transitions such as memory to calldata
Refinement-based optimization studies	Function-level gas reduction	Controlled contract experiments	Report on measurable runtime savings	Usually focus on specific constructs rather than combining storage-packing and calldata redesign
This study	Variable packing + calldata optimization	Private Geth v1.13.5 + debug_traceTransaction	40.68% saving for storeData(); 28.50% average saving	Provide deterministic opcode-level evidence linking savings to reduced storage intensity

TABLE VII. COST SAVINGS ACROSS TRANSACTION VOLUME AND GAS PRICE SCENARIOS

Transaction Volume (V)	Low (15 Gwei) ($USD)	Average (30 GWei) ($USD)	High (100 Gwei) ($USD)
1,000	1,605	3,210	10,700
2,500	4,013	8,025	26,750
5,000	8,025	16,050	53,500
7,500	12,038	24,075	80,250
10,000	16,050	32,100	107,000

Under the average scenario for gas prices (30 Gwei), this translates into an estimated saving of USD 3.21 per transaction. Although this is a small amount at the level of a single transaction, the accumulated savings are large when transaction volume grows. For an application with 10,000 annual transactions, the savings in a year of 30Gwei would be USD 32,100.

Optimized contracts minimize the user-side risk of increased transaction costs to users in the event of increasing gas prices. The relationship between the volume of transactions and the total accumulated cost savings is shown in Fig. 11.

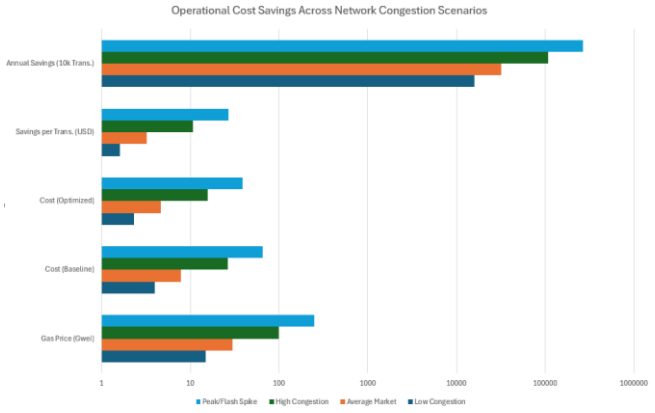


Fig. 11. Cost savings over transaction volume.

More critically, the effectiveness of this optimization increases linearly with both gas price and transaction volume. As can be seen from the sensitivity analysis, when the price of gas rises to 100 Gwei (High Scenario), the savings for the same volume of transactions rise to $107,000 USD. This means that code optimization can help reduce cost exposure when more expensive network conditions are present.

TABLE VIII. OPERATIONAL COST PER TRANSACTION UNDER DIFFERENT NETWORK SCENARIOS

Network Scenario	Gas Price (Gwei)	Cost (Baseline) ($USD)	Cost (Optimized) ($USD)	Savings per Transaction ($USD)	Annual Saving (10k Transaction) ($USD)
Low Congestion	15	3.95	2.34	1.61	16,050
Average Market	30	7.89	4.68	3.21	32,100
High Congestion	100	26.30	15.60	10.70	107,000
Peak / Flash Spike	250	65.75	39.00	26.75	267,500

This results in contract-level code efficiency that can complement the optimization strategies for Ethereum as a whole.

D. Optimization Paradox

Even though the empirical findings establish that smart contracts can offer quantifiable economic returns, optimization has structural trade-offs other than gas efficiency. Qualitative examination of the refactored source code highlights a practical trade-off: as gas efficiency improves, cognitive complexity and maintenance burden may also increase.

Techniques such as variable packing and bitwise substitution (SHL, OR) transform declarative and human-readable assignments into compact but semantically dense expressions. For example, a baseline implementation may be storing variables by direct assignments (e.g., a = x; b = y;). In contrast, an optimized implementation consolidates these values using explicit bit manipulation (e.g., data = (x << 128) | y;). While functionally equivalent, the optimized form requires deeper familiarity with low-level EVM semantics, increasing the cognitive load placed on developers and auditors.

Such a trade-off is mathematically described in Fig. 12, which describes the connection between the intensity of

optimization and its dual impact on the efficiency of gas consumption and the complexity of the cognitive process. The higher the intensity of optimization, the greater is the gas efficiency, which grows in something of a linear nature due to the reduced reliance on the expensive storage operations.

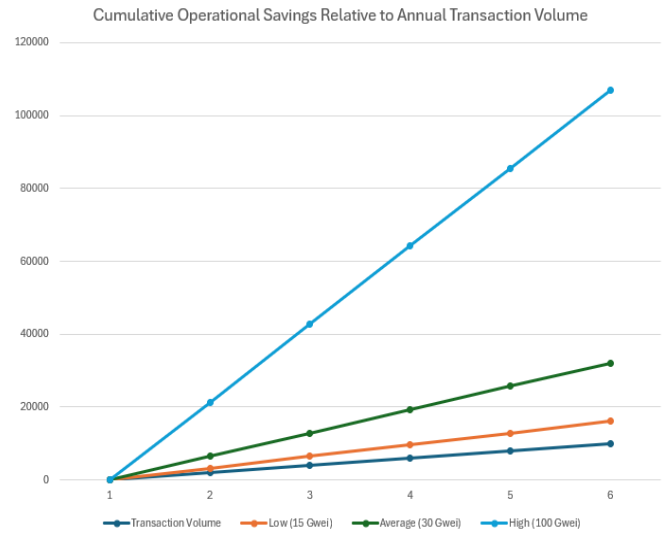


Fig. 12. Optimization paradox: trade-off between gas efficiency and cognitive complexity.

Nonetheless, there is an increasing possibility that cognitive complexity will increase disproportionately because of more intricate and less intuitive reasoning. Their difference indicates that there is a decision boundary that is critical, beyond which the positive effects of the marginal improvements in gas efficiency might be negated by the marginal increases in the risks involved with their usage.

Along with the aspects of readability, extreme optimization might introduce new security issues. Reducing the storage type (e.g., compressing uint256 to uint128 or uint64) would reduce storage use but limit the range of numbers. Though in the new versions of Solidity, checking for overflow is by default, given the desire to reduce the types intentionally, one still has to validate the range to avoid accidentally truncating or breaking the boundaries. To some degree, such safeguards can offset gas savings and make implementation more complicated.

The refactoring of the manual, therefore, should be done in a strategic rather than a universal way. Transactional settings with high frequency, including decentralized finance (DeFi) protocols or exchange mechanisms, can warrant more significant optimization because of gas savings accruing on large volumes of transactions. On the other hand, governance contracts, voting systems of DAOs, or those systems that prioritize transparency and auditability might not need to extend marginal efficiency benefits to readability.

The question of optimizing is hence a multi-dimensional cost-benefit analysis of user-side economic efficiency (User Experience) versus developer-side maintainability (Developer Experience). As implied by the balanced point in Fig. 12, it is not about maximizing smart contract engineering but about achieving the right balance between performance and structural clarity.

E. Validation Against EIP Standards

The empirical results of this study are in line with the economic incentives that are embedded in the Ethereum protocol, especially introduced by EIP-2028 and the block gas limit mechanism.

EIP-2028 reduced the gas cost of calldata to encourage off-chain data usage and minimize long-term state growth. The apparent savings due to the conversion from a memory-based parameter handling to the CALLDATA-based access reflect this protocol-level incentive structure. By moving data handling from more storage-intensive patterns or memory expanding patterns, the optimized implementation is generally consistent with Ethereum's disincentive for calldata that uses more storage or memory. The experimental trace results therefore show compatibility between the two levels of refactoring (opcode-level) and cost modeling (protocol-level).

Beyond individual transaction savings, optimization has implications at the block-level resource allocation layer. With a block gas limit of around 30 million units of gas, regarding the number of transactions that can be included in a block, the lower the per-transaction gas consumption is, the more of them can be included in a block. For example, if a contract's gas usage is decreased from 100,000 to 60,000 gas units the theoretical transaction capacity per block can be increased from around 300 transactions to 500 transactions under constant block limits. This doesn't quite double throughput, but it does significantly boost execution density without changing Layer-1 consensus parameters. These results illustrate that optimizing at the opcode level not only saves developers money, but it also makes better use of resources in the Ethereum limited execution environment.

Gas-aware architectural solutions don't just save money for each person; they also make assessed execution traces less storage-intensive and improve the use of blocks given the current protocol limits.

So, manual optimization should be thought of as a design method that fits inside Ethereum's economic structure and works with protocol-level scaling efforts instead of replacing them.

The results ought to be seen as data derived from controlled contract-level studies, rather than as direct validation of network-wide scalability, fee-market, or state-growth impacts inside the Ethereum ecosystem.

V. Conclusion

The effect of Solidity refactoring by hand was studied in terms of the consumption of gas in Ethereum smart contracts in a controlled Geth v1.13.5 environment for execution. By comparing baseline and optimized contract variants, the study quantified function-level gas savings, analyzed opcode-level execution changes, and estimated scenario-based cost implications.

The results indicate that manual refactoring by variable packing and calldata optimization has reduced gas consumption for the evaluated functions, with the greatest reduction in gas consumption of 40.68% for storage-intensive execution, while the average reduction in gas consumption for the tested functions is 28.5%. Opcode trace analysis further indicated that these savings were mainly associated with reduced reliance on

costly storage-write operations and greater use of lower-cost bitwise instructions. Scenario-based estimates also indicated that observed gas savings can lead to significant cost reductions with higher gas prices. Overall, the results indicate that smart contract optimization can benefit from a better understanding of EVM-level execution behavior rather than just source-level programming practice.

From a more practical standpoint, these results suggest that gas efficiency should be addressed together with correctness, readability, and maintainability while smart contracts are being designed. The evaluated refactoring pattern also aids storage-conscious contract engineering by optimizing the storage use within the contract design, which was tested. However, these findings should be interpreted as contract-level evidence under controlled conditions rather than as proof of ecosystem-wide scalability, fee-market, or state-growth effects.

There are a few limitations of this study. The first is that the experiments are only run with four contract functions and two types of optimization patterns: storage-slot packing and calldata-based parameter handling. Therefore, the findings should be interpreted as controlled evidence for CRUD-style, storage-heavy contracts rather than as general evidence for more complex architectures such as inheritance-based, upgradeable, DeFi-oriented, or cross-contract designs. Second, the refactoring was performed manually, which may increase developer cognitive load and reduce repeatability without automated support. Thirdly, the study did not employ formal verification and equivalence-checking tools; hence, semantic equivalence was not formally proven, but rather experimented with. Lastly, the analysis is primarily based on gas usage and opcode traces and does not fully measure maintainability, readability, auditability, test coverage, or the security of the optimized code.

Future studies could focus on testing more contract architectures, including Inheritance-based contracts, Upgradeable proxy contracts, Token standards, Access-control patterns, DeFi-oriented workflows, and Cross-contract interactions. Future studies should also develop automated or semi-automated refactoring support to reduce developer cognitive load and improve repeatability. Furthermore, formal verification, equivalence checking, static security analysis, test coverage analysis, and maintainability/readability measures should be used in conjunction with gas optimization to assess the savings in execution time, in addition to correctness, security, and the long-term quality of the code.

Overall, the results show that manual refactoring can reduce execution gas under controlled CRUD-style conditions, but broader adoption requires validation across more complex contract designs and complementary assessment of semantic equivalence, security, readability, and maintainability.

ACKNOWLEDGMENT

This research was supported by a UTM Fundamental Research Grant Q.J130000.3828.23H38.

REFERENCES

[1] D. D. Wood, “ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER,” 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:4836820

[2] B. Hu et al., “A comprehensive survey on smart contract construction and execution: paradigms, tools, and systems,” Patterns, vol. 2, no. 2, p. 100179, Feb. 2021, doi: 10.1016/j.patter.2020.100179.

[3] B. K. Meister and H. C. W. Price, “Gas fees on the Ethereum blockchain: from foundations to derivative valuations,” Front. Blockchain, vol. Volume 7-2024, 2024, doi: 10.3389/fbloc.2024.1462666.

[4] F. A. Almalki and A. Rajaram, “Optimizing gas efficiency and enhancing security in Ethereum smart contracts through integrated clustering and anomaly detection,” Int. J. Inf. Secur., vol. 24, no. 3, p. 145, May 2025, doi: 10.1007/s10207-025-01057-5.

[5] M. M. Khan, F. S. Khan, M. Nadeem, T. H. Khan, S. Haider, and D. Daas, “Scalability and Efficiency Analysis of Hyperledger Fabric and Private Ethereum in Smart Contract Execution,” Computers, vol. 14, no. 4, 2025, doi: 10.3390/computers14040132.

[6] S. Leonardos, B. Monnot, D. Reijsbergen, E. Skoulakis, and G. Piliouras, “Dynamical analysis of the EIP-1559 Ethereum fee market,” in Proceedings of the 3rd ACM Conference on Advances in Financial Technologies, in AFT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 114–126. doi: 10.1145/3479722.3480993.

[7] C. Li, “Gas Estimation and Optimization for Smart Contracts on Ethereum,” in 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2021, pp. 1082–1086. doi: 10.1109/ASE51524.2021.9678932.

[8] D. R. de Lima Cabral, P. Antonino, and A. C. A. Sampaio, “Demystification and near-perfect estimation of minimum gas limit and gas used for Ethereum smart contracts,” J. Cloud Comput., vol. 14, no. 1, p. 29, Jun. 2025, doi: 10.1186/s13677-025-00751-y.

[9] M. T. Ta and T. Q. Do, “A study on gas cost of ethereum smart contracts and performance of blockchain on simulation tool,” Peer--Peer Netw. Appl., vol. 17, no. 1, pp. 200–212, Jan. 2024, doi: 10.1007/s12083-023-01598-3.

[10] S. Park, J. Lee, and H. Kim, “Efficient computation offloading for ethereum DApps,” J. Ind. Inf. Integr., vol. 31, p. 100411, 2023, doi: https://doi.org/10.1016/j.jii.2022.100411.

[11] Q.-P. Kong et al., “Characterizing and Detecting Gas-Inefficient Patterns in Smart Contracts,” J. Comput. Sci. Technol., vol. 37, no. 1, pp. 67–82, Feb. 2022, doi: 10.1007/s11390-021-1674-4.

[12] T. Chen et al., “Towards saving money in using smart contracts,” in Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, in ICSE-NIER ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 81–84. doi: 10.1145/3183399.3183420.

[13] B. Gao, S. Shen, L. Shi, J. Li, J. Sun, and L. Bu, “Verification Assisted Gas Reduction for Smart Contracts,” in 2021 28th Asia-Pacific Software Engineering Conference (APSEC), 2021, pp. 264–274. doi: 10.1109/APSEC53868.2021.00034.

[14] K. Nelaturu, S. M. Beillahi, F. Long, and A. Veneris, “Smart Contracts Refinement for Gas Optimization,” in 2021 3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS), 2021, pp. 229–236. doi: 10.1109/BRAINS52497.2021.9569819.

[15] Q.-T. Nguyen, B. S. Do, T. T. Nguyen, and B.-L. Do, “GasSaver: A Tool for Solidity Smart Contract Optimization,” in Proceedings of the Fourth ACM International Symposium on Blockchain and Secure Critical Infrastructure, in BSCI ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 125–134. doi: 10.1145/3494106.3528683.

[16] M. Sipos and S. Szénási, “Optimal Gas Consumption in Ethereum Smart Contracts: A Targeted Review of Empirical Results, Design Patterns and Formal Methods,” in 2025 IEEE 25th International Symposium on Computational Intelligence and Informatics (CINTI), 2025, pp. 517–522. doi: 10.1109/CINTI67731.2025.11311839.

[17] A. D. Sorbo, S. Laudanna, A. Vacca, C. A. Visaggio, and G. Canfora, “Profiling gas consumption in solidity smart contracts,” J. Syst. Softw., vol. 186, p. 111193, 2022, doi: https://doi.org/10.1016/j.jss.2021.111193.

[18] E. Albert, J. Correas, P. Gordillo, G. Román-Díez, and A. Rubio, “GASOL: Gas Analysis and Optimization for Ethereum Smart Contracts,” in Tools and Algorithms for the Construction and Analysis of

Systems, A. Biere and D. Parker, Eds., Cham: Springer International Publishing, 2020, pp. 118–125.

[19] Shuwei S., Ni X., and Ting C., "Gas Optimization for Smart Contracts: A Survey," J. Comput. Res. Dev., vol. 60, no. 2, p. 311, 2023, doi: 10.7544/issn1000-1239.202220887.

[20] A. Kuhlman and A. Wicaksana, "Smart contract optimization for gas fee reduction with static solidity optimizer," Discov. Appl. Sci., vol. 7, no. 8, p. 888, Aug. 2025, doi: 10.1007/s42452-025-07537-6.

[21] M. He et al., "How to Save My Gas Fees: Understanding and Detecting Real-world Gas Issues in Solidity Programs." 2025. [Online]. Available: https://arxiv.org/abs/2403.02661

[22] J. Li, Z. Zhao, Z. Su, and W. Meng, "Gas-expensive patterns detection to optimize smart contracts," Appl. Soft Comput., vol. 145, p. 110542, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110542.

[23] T. Roughgarden, "Transaction Fee Mechanism Design for the Ethereum Blockchain: An Economic Analysis of EIP-1559." 2020. [Online]. Available: https://arxiv.org/abs/2012.00854

[24] M. Wijaya, F. Melvyn, R. Setiawan, and R. Y. Rumagit, "Ethereum vs Solana: A Comparative Study of Blockchain Architecture on Performance, Security, and Ecosystem Development," Procedia Comput. Sci., vol. 269, pp. 200–217, 2025, doi: https://doi.org/10.1016/j.procs.2025.08.273.

[25] B. Yu, T. Zhou, H. Zhao, X. Li, Y. Fan, and L. Chen, "Intra-node transaction parallelism in blockchains: Models, solutions, and trends," Comput. Sci. Rev., vol. 59, p. 100853, 2026, doi: https://doi.org/10.1016/j.cosrev.2025.100853.

[26] A. Jyoti, P. Gupta, S. Gupta, H. Khatter, and A. Mishra, "Inherent Insights using Systematic Analytics of Developments Tools in Ethereum Blockchain Smart Contract," Recent Adv. Electr. Electron. Eng., vol. 18, no. 2, pp. 135–146, 2025, doi: https://doi.org/10.2174/0123520965249434231024111732.

[27] P. S. Chakraborty and S. Tripathy, "SESIV: Secure and efficient smart contract based integrity verification of outsourced data," J. Inf. Secur. Appl., vol. 93, p. 104121, 2025, doi: https://doi.org/10.1016/j.jisa.2025.104121.

[28] E. Ni, E. Knight, and M. Gerstein, "Scalable and efficient on-chain data management in blockchain for large biomedical data," J. Biomed. Inform., vol. 165, p. 104818, 2025, doi: https://doi.org/10.1016/j.jbi.2025.104818.

[29] D. Marmsoler, "Deductive verification of solidity smart contracts with SSCalc," Sci. Comput. Program., vol. 243, p. 103267, 2025, doi: 10.1016/j.scico.2025.103267.

[30] J. P. Madrigal-Cianci, C. M. Maya, and L. Breakey, "A methodology for pricing gas options in blockchain protocols," Finance Res. Lett., vol. 84, p. 107700, 2025, doi: https://doi.org/10.1016/j.frl.2025.107700.

[31] N. Hejazi, "A Comprehensive Survey of Smart Contracts Vulnerability Detection Tools: Techniques and Methodologies," J. Netw. Comput. Appl., vol. 237, p. 104142, 2025, doi: 10.1016/j.jnca.2025.104142.

[32] A. Biere and D. Parker, Eds., Tools and Algorithms for the Construction and Analysis of Systems: 26th International Conference, TACAS 2020, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2020, Dublin, Ireland, April 25–30, 2020, Proceedings, Part II, vol. 12079. in Lecture Notes in Computer Science, vol. 12079. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-45237-7.

[33] J. Feist, G. Grieco, and A. Groce, "Slither: A Static Analysis Framework for Smart Contracts," in 2019 IEEE/ACM 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB), Montreal, QC, Canada: IEEE, May 2019, pp. 8–15. doi: 10.1109/WETSEB.2019.00008.

[34] W. Hughes, T. Magnusson, A. Russo, and G. Schneider, "Cheap and secure metatransactions on the blockchain using hash-based authorisation and preferred batchers," Blockchain Res. Appl., vol. 4, no. 2, p. 100125, Jun. 2023, doi: 10.1016/j.bcra.2022.100125.

[35] A. Liu, J. Chen, S. Yao, K. He, and R. Du, "An auditable and privacy-preserving user-controllable group signature scheme in blockchain," J. Inf. Secur. Appl., vol. 93, p. 104181, Sep. 2025, doi: 10.1016/j.jisa.2025.104181.

[36] F. R. Vidal, N. Ivaki, and N. Laranjeiro, "Analyzing the impact of elusive faults on blockchain reliability," Blockchain Res. Appl., vol. 6, no. 4, p. 100295, Dec. 2025, doi: 10.1016/j.bcra.2025.100295.

[37] N. Afraz, F. Wilhelmi, H. Ahmadi, and M. Ruffini, "Blockchain and Smart Contracts for Telecommunications: Requirements vs. Cost Analysis," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3309423.

[38] E. Albert, J. Correas, P. Gordillo, G. Román-Díez, and A. Rubio, "GASOL: Gas Analysis and Optimization for Ethereum Smart Contracts," in Tools and Algorithms for the Construction and Analysis of Systems, A. Biere and D. Parker, Eds., Cham: Springer International Publishing, 2020, pp. 118–125.

[39] T.-D. Tran, Q. Vu, B. Huynh, and V.-H. Pham, "Acheron: a market-based multi-relay architecture for adaptive and secure cross-chain communication," Internet Things, vol. 35, p. 101836, 2026, doi: https://doi.org/10.1016/j.iot.2025.101836.

[40] N. Andola and V. K. Yadav, "A blockchain-assisted privacy-preserving framework for Mobile CrowdSensing," Pervasive Mob. Comput., vol. 115, p. 102125, 2026, doi: https://doi.org/10.1016/j.pmcj.2025.102125.

[41] S.-C. Susan, "Leveraging Slither and Interval Analysis to build a Static Analysis Tool," Electron. Proc. Theor. Comput. Sci., vol. 410, pp. 150–166, Oct. 2024, doi: 10.4204/eptcs.410.10.

[42] H. Li, G. Xiong, C. Hou, G. Gou, Z. Chen, and Z. Li, "Smart Contract Vulnerability Detection Based on AST-Augmented Heterogeneous Graphs," in 2024 IEEE International Performance, Computing, and Communications Conference (IPCCC), 2024, pp. 1–10. doi: 10.1109/IPCCC59868.2024.10850061.

Agent-Oriented Fuzzy Decision Support System for Multi-Criteria Evaluation of Sustainable Investments in the Agro-Industrial Sector

Aigerim Omurtayeva[1], Ulzhan Makhazhanova[2]*, Akmaral Kulamanova[3],
Dinara Kargabaeva[4], Bolat Tassuov[5]*, Adilbek Tanirbergenov[6]*
Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan [1, 2]
Secondary School No.17 Named after Gani Muratbayev, Department of Education of Taraz City, Zhambyl Region, Kazakhstan [3]
Department of Information Systems, M.Kh. Dulaty Taraz University, Taraz, Kazakhstan[4]
Department of Physics and Computer Science, M.Kh. Dulaty Taraz University, Taraz, Kazakhstan[5]
Department of Cryptology, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan [6]

***Abstract*—Sustainable development of the agri-food sector in emerging economies requires the use of analytical tools capable of taking into account climate risks, environmental constraints, and investment flow instability when making management decisions. Given the fragmentary nature of statistical information and the high volatility of the external environment, traditional econometric methods for assessing investment attractiveness demonstrate limited effectiveness and low interpretability. This study proposes an agent-oriented modular fuzzy decision support framework for the comprehensive assessment of sustainable investments in the agricultural sector. The developed approach combines a modular data processing architecture that provides automated collection and preprocessing of heterogeneous statistical sources (OECD, FAO, and national statistics), with a fuzzy additive aggregation (Fuzzy-SAW) mechanism that allows for interpretable multi-criteria assessment of economic, environmental, and production-forecasting factors. The methodological novelty of the study lies in the integration of an automated data processing pipeline with an explainable fuzzy multi-criteria assessment model focused on conditions of data incompleteness and structural uncertainty. Empirical validation of the model was performed using statistical data from the agro-industrial complex of the Republic of Kazakhstan for the period 2010–2023. The results show that the proposed framework effectively smooths out short-term volatility in indicators and identifies long-term structural trends in investment attractiveness. In particular, in 2021–2023, the integral index of sustainable investment remained at around 0.37, despite adverse climate shocks, mainly due to the compensatory effect of growth in private investment flows, which indicates the formation of mechanisms for the adaptive sustainability of the agricultural sector. The proposed analytical framework is a scalable and interpretable decision support tool that can be used by government agencies, investors, and industry analysts in developing long-term strategies for sustainable agricultural development in emerging economies.**

Keywords*—Sustainable investments; agro-industrial sector; decision support system; agent-oriented systems; fuzzy logic; ESG factors; climate risks; emerging economies; multicriteria evaluation

I. Introduction

The intensification of climate risks, tightening environmental regulations, and high volatility of investment flows create a growing need for intelligent digital tools capable of supporting strategic decision-making through the integration of heterogeneous statistical data. Modern systems for assessing the sustainability of agricultural processes rely on data from multiple sources, including FAOSTAT, OECD, and national statistical databases, which differ significantly in structure, level of detail, and update frequency [1–3]. Problems of data interoperability, aggravated by external uncertainty and fluctuations in global markets, substantially limit the applicability of traditional centralized analytical models in forming long-term investment assessments.

The digital transformation of the agricultural sector is closely linked to the Agri-Food 4.0 paradigm, considered as an adaptation of Industry 4.0 principles for the integration of intelligent technologies, data interoperability, and real-time analytics [4]. The development of this concept has led to the emergence of Agricultural Digital Twins, used to model production processes and enhance the adaptive resilience of agricultural systems [5-7]. At the same time, artificial intelligence and data analytics methods are becoming key components of modern decision support systems; however, their practical effectiveness largely depends on data governance quality and the ability to integrate distributed information sources [8]. In this context, agent-oriented data-processing architectures are viewed as a promising computational mechanism for integrating distributed information flows, ensuring scalable processing of heterogeneous resources, and supporting managerial decision-making in complex dynamic environments [9].

Despite the rapid development of digital technologies, deterministic financial models still dominate many applied investment assessment tasks. These models are insufficiently adapted to the nonlinear dynamics of agricultural processes and the high degree of uncertainty associated with ESG factors [10–12]. The need to account for Environmental, Social, and Governance (ESG) factors significantly complicates investment analysis procedures, as the integration of economic, environmental, and forecast indicators requires the use of multicriteria computational methods [14–16]. Contemporary research shows that digital approaches to sustainable investment

*Corresponding author.

assessment are developing mainly along two trajectories: econometric models with limited capacity to incorporate expert uncertainty, and machine learning methods that provide high computational accuracy but lack sufficient interpretability. This contradiction creates a methodological gap between analytical performance and the practical applicability of models in decision support systems.

The use of fuzzy logic and multicriteria analysis methods partially addresses this gap by formalizing qualitative expert judgments and integrating heterogeneous indicators into interpretable sustainability indices [17–19]. However, existing studies typically consider multicriteria analysis methods and data integration architectures separately, limiting the potential for building scalable computational systems for strategic investment analysis. As a result, there remains a shortage of integrated computational architectures that combine agent-oriented data engineering mechanisms with interpretable fuzzy multicriteria models for sustainable investment assessment.

This study aims to bridge this methodological gap by developing an integrated hybrid Agent-Oriented Data Integration – Fuzzy-MCDM architecture that ensures automated integration of distributed international statistical data, interpretable multicriteria evaluation of sustainable investments, and robustness testing of the investment index based on scenario sensitivity analysis. The objective of the research is to design and empirically validate an intelligent decision support system that combines an agent-oriented data integration architecture with fuzzy logic for multicriteria evaluation of sustainable investments in the agro-industrial sector.

The main contribution of the study lies in: (1) the development of a modular agent-oriented architecture for integrating heterogeneous data sources (OECD, FAO, and national statistics); (2) the formulation of an interpretable fuzzy MCDM framework based on Fuzzy SAW with AHP-derived weighting coefficients; and (3) the empirical validation of the proposed system using retrospective data from the agro-industrial sector of Kazakhstan for the period 2010–2023. The results confirm the computational robustness of the model and its applicability to strategic investment analysis in emerging agricultural economies. Although Kazakhstan's data are used as an empirical case, the proposed approach is methodologically universal and can be adapted to national statistical databases of other countries characterized by high market volatility and fragmented statistical information [20,21].

II. Materials and Methods

A. Data Sources and Preprocessing

Within the proposed agent-oriented modular architecture, input data are treated as heterogeneous time series streams differing in update frequency (quarterly and annual reports) and measurement units. To ensure data integrity and suitability for fuzzy inference procedures, a preprocessing pipeline was implemented, including three sequential stages: temporal alignment, missing value imputation, and outlier filtering.

1) Temporal alignment: Since the initial indicators are provided with different temporal granularities (for example, foreign direct investment data are published quarterly, whereas OECD producer support indicators are annual), aggregation to a unified annual interval was performed. For additive indicators, quarterly values were aggregated by summation:

$$X_{\text{annual}}^{(t)} = \sum_{q=1}^{4} x_{\text{quarter}}^{(t,q)}$$

where, $x_{quarter}^{(t,q)}$ denotes the value of the indicator in the q quarter of year t. For point-in-time indicators (price indices, exchange rates), a weighted average value is applied. This approach eliminates temporal lags and ensures consistency in the dynamics of indicators.

2) Data imputation. In the historical time series, the proportion of missing observations did not exceed 5% of the total sample size. Linear interpolation was applied to restore missing values:

$$x_t = x_{t-1} + \frac{x_{t+1} - x_{t-1}}{2}$$

The choice of this method is justified by the monotonic dynamics of the indicators over short annual intervals, which helps minimize artificial volatility. Linear interpolation provides acceptable accuracy for time series with low variability, reducing bias compared to mean-value imputation [22].

Outlier Detection. To identify anomalies caused by technical aggregation errors or isolated shock events, a statistical filter based on the Z-score was applied. An observation x_t is classified as an outlier if the following condition holds:

$$|Z_t| = \left|\frac{x_t - \mu}{\sigma}\right| > 3$$

where, μ is the moving average and σ is the standard deviation over the analyzed period (the 3σ rule);

where, μ and σ denote the moving average and standard deviation, respectively (the 3σ rule). If an anomaly is detected, the value is replaced with the threshold $\mu \pm 3\sigma$, which prevents distortion of the integral index by extreme outliers. The design of this module was based on widely accepted principles of data validation in distributed analytical systems, aimed at ensuring reproducibility and computational robustness [23].

B. Formalization of the Input Indicator Space

To construct the fuzzy multicriteria model, the feature space is formalized as a vector of criteria $C = \{C_{PSE}, C_{AEI}, C_{Outlook}, C_{FDI}$, reflecting the economic, environmental, and forecast components of sustainable investments in the agro-industrial sector.

The selection of criteria is motivated by the need for a comprehensive consideration of ESG factors and macroeconomic dynamics:

1) Economic Support C_{PSE}: This criterion is based on the OECD Producer Support Estimate (PSE) [1], which aggregates government subsidies, price support, and transfers to producers. In the model, this indicator is treated as a *benefit-type* criterion, since higher institutional support reduces operational risks and enhances investment sustainability.

2) Environmental Load C_{AEI} : Formed based on Agri-Environmental Indicators (AEI), including mineral fertilizer usage intensity and greenhouse gas emissions. Following the ESG approach, this criterion is treated as a cost-type indicator, reflecting long-term environmental risks and potential constraints on sustainable development [24].

3) Market Potential C_{PSO}: This indicator is derived from OECD–FAO Agricultural Outlook forecast data [2] and is used as a benefit criterion, adjusting the current assessment based on expected medium-term production and consumption trends (5–10-year horizon).

4) Investment Climate C_{FDI}: Characterized by inflows of foreign direct investment (FDI) and used as a proxy for macroeconomic stability and external market confidence [25].

The combination of these criteria forms an interpretable feature space, enabling the integration of economic, environmental, and forecast indicators within a single fuzzy framework for supporting sustainable investment decisions.

To ensure comparability of indicators expressed in different physical units (USD million, thousand tons of CO_2-equivalent, indices), Min–Max normalization was applied, scaling all values to the interval [0,1]. Transformations were applied according to the criterion type:

For *benefit-type* criteria (C_{PSE}, C_{PSO}), standard linear normalization was used:

$$\mu_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

where, x_{ij} is the value of the j-th indicator for the i -th observation year.

For the investment climate indicator (C_{AEI}), considering saturation effects:

$$\mu_{FDI}(x) = \min\left(1, \frac{x}{Target_{FDI}}\right)$$

where, x is the investment volume, and $Target_{FDI} = 300$ million USD represents the empirically determined saturation level corresponding to a stable capital inflow into the agro-industrial sector.

For the environmental criterion C_{AEI}, a cost-type indicator, a linearly decreasing function was applied:

$$\mu_{AEI}(x) = \max\left(0, 1 - \frac{x}{Limit_{CO2}}\right)$$

where, x is the emission volume, and $Limit_{CO2} = 50$ million tons corresponds to the sector's (assimilative capacity) and serves as a critical threshold for forming the sustainable investment index.

The application of these transformations allows heterogeneous indicators to be brought into a unified normalized space, ensuring correct integration of economic, environmental, and forecast factors within the fuzzy multi-criteria analysis of sustainable investments.

C. Implementation of Agent-Oriented Modular Architecture

To automate the processes of data collection, harmonization, and analytical processing, an agent-oriented modular computing architecture was implemented in the study. Unlike classical multi-agent systems focused on modeling the interaction of autonomous actors, in this work, agents are interpreted as specialized software modules that perform formalized operations of data extraction, transformation, and preparation for subsequent fuzzy multi-criteria analysis.

This approach allows us to preserve the principles of agent-oriented design (modularity, function decomposition, extensibility) while avoiding assumptions about complex mechanisms of negotiation or competition between agents, which makes the architecture more transparent and reproducible for sustainable investment analysis tasks.

The computational implementation of the system was done in Python 3.9. The pandas library was used for time series processing, data aggregation, and preprocessing operations. Fuzzy normalization and membership function formation procedures were implemented using the scikit-fuzzy library (v0.4.2).

To improve the efficiency of working with distributed statistical sources, an asynchronous input-output model based on the asyncio library was used, which allowed data to be loaded without blocking the main computing thread. Interaction between software modules is organized through the transfer of structured data in JSON format, ensuring sequential activation of processing and analysis stages.

The system architecture is organized according to a hierarchical principle and includes three functional levels that ensure sequential data integration, fuzzy processing, and aggregation of results.

Level 1 — Data Source Modules: This level implements specialized software modules responsible for extracting data from external sources (OECD, FAO, national statistics). Each module encapsulates the logic for accessing the corresponding source, including data availability checks, basic structural validation of input files (CSV/JSON), and conversion of data to a unified format.

Level 2 — Analytical Modules: At this stage, preprocessing and fuzzification of normalized indicators are performed. Clear quantitative values are converted into linguistic assessments ("low", "medium", "high") using symmetric triangular membership functions. This level provides a transition from numerical time series to interpretable fuzzy representations used in further multi-criteria analysis.

Level 3 — Aggregation Module: The final level implements the fuzzy additive weighting (Fuzzy SAW) procedure, aggregating membership degrees across all criteria and forming an integral index of sustainable investments. The weight coefficients of the criteria are determined by the analytic hierarchy process (AHP), and the consistency of expert assessments is controlled using the consistency coefficient.

The proposed agent-oriented modular architecture provides a reproducible sequence of computational steps, including data extraction, normalization, fuzzification, and aggregation of indicators. This organization of the computational process increases the transparency of the integral index formation and makes the system suitable for practical use in sustainable investment assessment tasks.

A schematic representation of data flows and functional levels of the architecture is shown in Fig. 1.

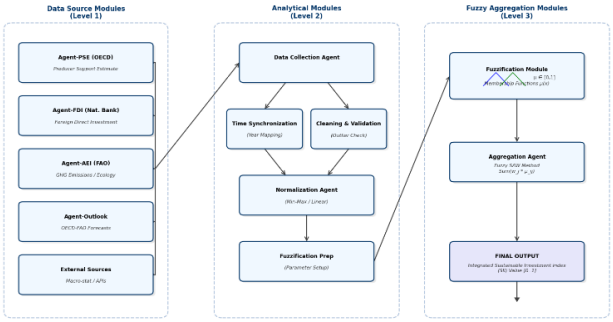


Fig. 1. Modular agent-oriented architecture for data collection, preprocessing, and fuzzy aggregation.

Despite the use of well-known methods of fuzzy logic, multi-criteria analysis, and agent-oriented data processing, the methodological novelty of this study lies in their structured integration into a unified decision support framework for assessing sustainable investments. Unlike existing approaches, in which the stages of data integration, normalization, and evaluation are considered separately, the proposed architecture formalizes the relationship between input data characteristics, membership function selection, and aggregation mechanism. This ensures consistency in computational procedures, improves the interpretability of results, and enables robust evaluation in conditions of data heterogeneity and uncertainty. Thus, the contribution of the study lies not in the development of a new fuzzy operator, but in the formalization of a reproducible integrated computational process focused on the tasks of robust investment analysis.

D. Model Validation Strategy

To verify the stability and correctness of the proposed model, a multi-level validation procedure was used, including:

1) Sensitivity analysis — analysis of the sensitivity of the integral index to variations in the weighting coefficients of the criteria (±10–20%), allowing the stability of the investment attractiveness ranking to be assessed.

2) Temporal consistency check — verification of the consistency of the index dynamics with macroeconomic and climatic events of the period under study.

3) Comparative robustness assessment — comparison of the model results with trends in key industry indicators (investment, government support, production dynamics), which allows confirming the interpretability of the obtained estimates.

The application of these procedures ensures the robustness of the model and confirms the reproducibility of the results obtained when the input parameters change.

III. Results

A. Normalization and Fuzzy Representation of Input Data

In accordance with the proposed agent-oriented modular architecture of the hybrid fuzzy MCDM framework, the first stage of the computational experiment involved converting heterogeneous statistical indicators to a unified dimensionless interval [0,1], which ensured the comparability of economic, environmental, and forecast variables.

To fuzzify the normalized data, piecewise linear membership functions (triangular and trapezoidal, formulas 1–4) were used, providing an interpretable mapping of quantitative values to linguistic categories. This choice allows for transparent tracking of the impact of changes in individual indicators (e.g., the level of government support or the volume of investment) on the integral index of sustainable investment.

Let x_i denote the aggregate value formed by data agents. The boundaries of the universe $[x_{min}, x_{max}]$ were set with an expansion±15% relative to historical extremes for the period 2010–2023, which allowed taking into account potential fluctuations in input indicators when modeling scenarios and avoiding values going beyond the domain of membership functions.

1) Assessment of institutional support for producers C_{PSE}: The Producer Support Estimate (PSE) indicator characterizes the level of transfers from consumers and taxpayers to agricultural producers. According to OECD data, institutional support for Kazakhstan's agricultural sector in 2010–2023 varied significantly, taking both negative values (reflecting the fiscal burden on the sector) and positive levels reaching 4–5% of gross farm income in certain years.

Within the model, the indicator C_{PSE} is interpreted as a benefit-type criterion and normalized using an increasing linear membership function:

$$\mu_{PSE}(x) = \begin{cases} 1, & x \le x_{min} \\ \frac{x - x_{min}}{x_{max} - x_{min}}, & x_{min} < x < x_{max} \\ 0, & x \ge x_{max} \end{cases} \quad (1)$$

where, x_{max} is set at 15%, which exceeds the historical maximum (13.21% in 2012) and allows for the potential expansion of government support programs.

Applying this membership function to the 2010–2023-time series (Table III) made it possible to quantitatively characterize the dynamics of the institutional environment. In particular, in 2012, the maximum value of $\mu_{PSE} = 0{,}881$ was recorded, reflecting the peak of state support. In 2016, the negative value of PSE (-1.70%) was converted to $\mu_{PSE} = 0$, which is interpreted by the model as a lack of institutional stimulation of the sector during this period. In 2021 – 2022, there was a decline in normalized values, coinciding with the introduction of export restrictions and anti-inflationary measures, followed by a recovery in the indicator in 2023.

As an illustration, let us take the calculation for 2015. According to OECD data, the PSE value $x = 11{,}85\%$..With $x_{min} = 0$ and $x_{max} = 15$, we get:

$$\mu_{PSE}(11{,}85) = \frac{11{,}85}{15} = 0{,}79$$

The resulting value reflects the high level of institutional support for the agricultural sector in the year under review.

Table I presents the results of normalizing the C_{PSE} indicator for the period 2010–2023, demonstrating the cyclical nature of government support and its contribution to the formation of the integral index of sustainable investments.

TABLE I. NORMALIZED VALUES OF THE INSTITUTIONAL SUPPORT INDICATOR FOR THE AGRICULTURAL SECTOR C_{PSE} FOR 2010–2023 (DATA SOURCE: OECD.STAT)

Year	PSE value (%)*	Calculation using formula (1)	μ_{PSE}	Level of support
2010	8.52	8.52/15	0.568	Medium-High
2011	11.34	11.34/15	0.756	High
2012	13.21	13.21/15	0.881	High
2013	12.25	12.25/15	0.817	High
2014	4.84	4.84/15	0.323	Low-Medium
2015	11.85	11.85/15	0.790	High
2016	-1.70	<0	0.000	None
2017	5.65	5.65/15	0.377	Low-Medium
2018	5.22	5.22/15	0.348	Low-Medium
2019	10.10	10.10/15	0.673	Medium-High
2020	6.81	6.81/15	0.454	Medium
2021	7.02	7.02/15	0.468	Medium
2022	4.49	4.49/15	0.299	Low
2023	9.19	9.19/15	0.613	Medium-High

[a.] A negative PSE value in 2016 indicates fiscal pressure on producers (e.g., through export restrictions) and is interpreted in the model as a lack of stimulus effect. The threshold of x_max=15% was selected based on the historical maximum and is used as the upper limit for normalization.

2) Foreign direct investment inflows C_{FDI}: The foreign direct investment (FDI) inflows indicator is used in the model as a proxy metric for the external capitalization of the agricultural sector. During the period under review, Kazakhstan's agriculture was characterized by significantly lower volumes of capital attracted compared to the raw materials sectors of the economy. According to data from the National Bank of the Republic of Kazakhstan, even in the year of maximum inflows (2015), the volume of investment amounted to US$71.8 million, while in most years the figures ranged from US$10 million to US$30 million.

Taking into account the effect of diminishing marginal utility of capital, a saturation function is used to normalize the indicator:

$$\mu_{FDI}(x) = \min(1, \frac{x}{Target_{FDI}}) \quad (2)$$

where, $Target_{FDI} = 300$ million is set as the indicative level of sustainable capitalization of the agricultural sector. This threshold exceeds historical values and is used as a long-term benchmark reflecting the structural need of the industry for investment resources.

The normalized values of μ_{FDI} are significantly lower than the corresponding values of institutional support μ_{PSE}, reflecting the continuing gap between the volume of state support and the external capitalization of the sector. This is consistent with the fact that the share of agriculture in the total inflow of FDI in Kazakhstan traditionally remains below 1%.

TABLE II. NORMALIZED VALUES OF THE EXTERNAL CAPITALIZATION INDICATOR FOR THE AGRICULTURAL SECTOR C_{FDI} FOR 2010–2023 (DATA FROM THE NATIONAL BANK OF KAZAKHSTAN)

Year	Gross FDI inflow, million USD*	Calculation according to formula (2)	μ_{FDI}	Level of investment activity
2010	6.0	6.0/300	0.02	Very Low
2011	7.8	7.8/300	0.026	Very Low
2012	18.3	18.3/300	0.061	Low
2013	5.3	5.3/300	0.018	Very Low
2014	-1.7	<0	0.000	None
2015	71.8	71.8/300	0.239	Low-Medium
2016	50.1	50.1/300	0.167	Low
2017	27.5	27.5/300	0.092	Low
2018	13.7	13.7/300	0.046	Very Low
2019	14.9	14.9/300	0.050	Very Low
2020	9.5	9.5/300	0.032	Very Low
2021	36.3	36.3/300	0.121	Low
2022	31.1	31.1/300	0.104	Low
2023	49.5	49.5/300	0.165	Low

Table II presents the results of normalizing the C_FDI indicator for the period 2010–2023. Even in 2015, which was characterized by the maximum capital inflow, the membership function value was only $\mu_{FDI} = 0{,}239,$ indicating a significant deviation from the saturation level. In the post-pandemic period, a gradual recovery of the indicator is observed (growth from 0.032 in 2020 to 0.165 in 2023), but the absolute values remain in the range of low and moderate investment activity.

The results obtained indicate that, within the framework of the model, the FDI acts primarily as an indicator of the sector's long-term potential rather than a factor of current sustainability, highlighting the structural undercapitalization of the agricultural economy.

3) Agro-ecological indicators C_{AEI}: The environmental component is based on FAOSTAT data on greenhouse gas emissions. For the agricultural sector of Kazakhstan, this indicator is about 29 – 32 million tons of CO_2 equivalent per year (about 9% of national emissions).

The environmental component of the model is based on FAOSTAT data on greenhouse gas emissions from the agricultural sector (Emissions Total, CO_2-eq). During the period under review (2010 – 2023), emissions ranged from 21 to 30.5 million tons of CO_2 -eq, accounting for about 9% of national emissions.

Within the model, the C_{AEI} indicator is considered a cost criterion: an increase in emissions corresponds to a decrease in the sector's environmental sustainability. A linear inverse membership function is used for normalization:

$$\mu_{AEI}(x) = \max(0{,}1 - \frac{x}{Limit_{CO2}}) \quad (3)$$

where, x is the current emission volume (million tons of CO_2-equivalent) according to FAO data;

$Limit_{CO2} = 50$ million tons is the threshold value of environmental (assimilative) capacity used as a benchmark in the normalization procedure.

TABLE III. NORMALIZED VALUES OF AGRO-ENVIRONMENTAL INDICATORS C_{AEI} FOR 2010–2023 (FAOSTAT)

Year	Emissions, million tons CO_2eq	Calculation according to formula (3)	C_{AEI}	Level of environmental impact
2010	26.34	1−0.527	0.473	Medium
2011	23.22	1−0.464	0.535	Medium–High
2012	22.88	1−0.458	0.542	High
2013	20.98	1−0.420	0.580	High
2014	23.75	1−0.475	0.525	Medium
2015	25.41	1−0.508	0.492	Medium–Low
2016	23.72	1−0.474	0.526	Medium
2017	30.47	1−0.609	0.391	Low
2018	26.18	1−0.524	0.476	Medium
2019	25.60	1−0.512	0.488	Medium
2020	24.69	1−0.494	0.506	Medium–High
2021	25.86	1−0.517	0.483	Medium
2022	24.43	1−0.489	0.511	Medium–High
2023	23.95	1−0.479	0.521	Medium–High

The normalization results (Table III) show moderate variability in the sector's environmental characteristics during the period analyzed. The maximum value of the membership function is observed in 2013 ($\mu_{AEI} = 0.580$), corresponding to the minimum level of emissions (20.98 million tons), while in 2017 the minimum value is recorded ($\mu_{AEI} = 0.391$), reflecting an increase in emissions to 30.47 million tons.

In the period 2020–2023, the indicator values stabilize in the range $\mu_{AEI} = 0.48 - 0.52$, which indicates the relative stability of the environmental impact of the agricultural sector amid changes in production volumes.

4) Production stability and prospects (C_{PSO}): The production stability indicator C_{PSO} is a separate component of the model that reflects the production stability of the agricultural sector and is based on the Physical Volume Index (PVI) published by the National Statistics Bureau of the Republic of Kazakhstan. A retrospective analysis of data for the period 2010–2023 shows significant variability in agricultural production, primarily due to climatic factors, including drought years.

The physical volume index (previous year=100%) is used to formalize the indicator. In order to correctly account for deviations from the level of stable production, a sigmoidal membership function is applied:

$$\mu_{\text{PSO}}(x) = \frac{1}{1+e^{-0{,}2\cdot(x-100)}} \quad (4)$$

where, x is the value of the Physical Volume Index (PVI), %; 100 is the baseline threshold (stable production level); $\alpha = -0{,}2$ is the sensitivity coefficient characterizing the elasticity of the sector.

TABLE IV. NORMALIZED VALUES OF THE PRODUCTION SUSTAINABILITY INDICATOR C_{PSO} BASED ON PVI DATA (2010–2023)

Year	PVI (%) (Official Data)	μ_{PSO} (Fuzzy assessment)	Production stability level
2010	89.6 (Drought)	0.111	Very Low
2011	121.4 (Record)	0.987	High
2012	85.3 (Decline)	0.050	Very Low
2013	109.6	0.872	High
2014	101.0	0.550	Medium
2015	103.4	0.664	Medium–High
2016	105.4	0.746	High
2017	103.1	0.650	Medium–High
2018	103.5	0.668	Medium–High
2019	99.9	0.495	Medium
2020	105.7	0.758	High
2021	97.7 (Drought)	0.387	Low–Medium
2022	109.1	0.860	Medium–High
2023	91.6 (Drought)	0.157	Very Low

The normalization results (Table IV) show alternating periods of increased and decreased production stability. The maximum values of the membership function are observed in years of significant increase in output (e.g., 2011, 2013, and 2022), while the minimum values correspond to drought years characterized by a decline in the production index. In particular, in 2023, the PVI value was 91.6%, which corresponds to $\mu_{\text{PSO}} = 0.157$,, reflecting the reduced production stability of the sector.

Overall, the results confirm the high sensitivity of agricultural production to climatic fluctuations, which justifies the need to take this indicator into account in the integral assessment of sustainable investments.

B. The Process of Fuzzy Aggregation and the Formation of an Integral Index of Sustainable Investments (Hybrid Fuzzy-AHP Approach)

After the fuzzification stage, the state space of the system is represented by a vector of normalized membership degrees $\mu_{ij}(x_i)$, where each component reflects the degree of

favorability j of the factor (economic, environmental, or prognostic) for investments in the period t.

After the fuzzy faction stage, the state of the system is represented by a vector of normalized membership degrees $\mu_{ij}(x_i)$, where each component reflects the relative state of economic, environmental, and production factors in the period t.

To integrate private criteria, a hybrid Fuzzy SAW (Fuzzy Simple Additive Weighting) approach is used, which combines the interpretability of fuzzy estimates with formal multi-criteria aggregation. This approach ensures transparency of calculations and comparability of the impact of various sustainable development factors.

1) Determination of weighting coefficients (AHP): The relative significance of the criteria ω_j was determined using the Analytic Hierarchy Process (AHP) based on expert assessments reflecting the balance between institutional, investment, environmental, and production factors. To verify the consistency of expert judgments, a consistency ratio was calculated. The resulting value (CR=0.067) was below the threshold level of 0.10, which confirms the correctness of the expert assessments and the reliability of the weighting coefficients (Table V).

The weight coefficients satisfy the normalization condition:

TABLE V. MATRIX OF PAIRWISE COMPARISONS OF CRITERIA AND AHP WEIGHTS

Criterion	C_{PSE} (Support)	C_{FDI} (Investments)	C_{AEI} (Ecology)	$C_{Outlook}$ (Forecast)	Weight (ω_j)
C_{PSE}	1	2	4	3	0.46
C_{FDI}	1/2	1	3	2	0.28
C_{AEI}	1/4	1/3	1	1/2	0.09
$C_{Outlook}$	1/3	1/2	2	1	0.17
Total	02.08	3.83	10.0	6.5	1.00

The matrix is constructed on the Saati scale (1–9). The consistency ratio was CR = 0.067, which is below the threshold value of 0.10 and confirms an acceptable level of consistency in expert assessments.

The weight coefficients satisfy the normalization condition:

$$\sum_{j=1}^{n} \omega_j = 1, \quad \omega_j \geq 0 \tag{5}$$

SII_t the integral index of sustainable investments for the period t is defined as the weighted sum of normalized fuzzy estimates:

$$SII_t = \sum_{j=1}^{n} \omega_j \cdot \mu_{ij}(x_{jt}) \tag{6}$$

where, ω_j are the weights of the criteria obtained by the AHP method, $\mu_{ij}(x_{jt})$ are the normalized degrees of membership of the corresponding indicators.

The obtained value $SII_t \in [0,1]$ reflects the overall state of stability of the investment environment in the agricultural sector, taking into account institutional, environmental, production, and investment factors.

The additive structure of the Fuzzy SAW method ensures the decomposability of the result, allowing for a quantitative assessment of the contribution of each factor to the formation of the integral index and analysis of the sensitivity of the final assessment to changes in individual criteria.

C. Dynamics of the Integral Index of Sustainable Investments SII and Interpretation of Structural Shifts

To identify sustainable long-term trends and eliminate high-frequency noise characteristic of commodity markets, the analysis of the SII dynamics was based on aggregated time intervals. The time series for 2010–2023 was discretized into four phases reflecting different stages of institutional and economic development of Kazakhstan's agro-industrial complex.

The calculation results obtained using the developed Fuzzy SAW model on real statistical data are presented in Table VI.

TABLE VI. AGGREGATED VALUES OF FUZZY ESTIMATES OF CRITERIA AND THE INTEGRAL INDEX SII BY PERIOD

Period	PSE (Support)	FDI (Investments)	AEI (Ecology)	Outlook (Forecast)	SII (Index)	Interpretation
2010 – 2013	0.76	0.03	0.53	0.51	0.49	High
2014 – 2017	0.37	0.13	0.48	0.65	0.36	Moderate
2018 – 2020	0.49	0.04	0.49	0.64	0.39	Stable
2021 – 2023	0.46	0.13	0.51	0.47	0.37	Mixed

The sharp decline in the PSE in the last phase is due to price containment measures (negative support according to OECD methodology), which temporarily lowered the overall index despite record growth in market factors (FDI and Outlook).

Phase analysis of the dynamics of the Sustainable Investment Index (SII):

Phase 1: Period of government momentum and institutional asymmetry (2010 – 2013). This period is characterized by the highest value of the $SII \approx 0.49$. Decomposition of the index reveals a pronounced institutional asymmetry of sustainability factors: a high level of government support ($\mu_{PSE} \approx 0.76$, peaking in 2012) compensated for limited private investment activity ($\mu_{FDI} \approx 0.03$). The formation of investment sustainability during this period was mainly due to government support measures in relatively favorable environmental conditions ($\mu_{AEI} \approx 0.53$).

Phase 2: Market correction and capacity building (2014 – 2017). During this period, the integral index of sustainable investment fell to $SII \approx 0.36$, corresponding to the moderate sustainability zone. The model recorded a structural shift in the system of factors: a reduction in the fuzzy assessment of state support ($\mu_{PSE} \approx 0.37$) was accompanied by the first statistically significant inflow of private foreign investment ($\mu_{FDI} \approx 0.13$). This dynamic reflects the beginning of the transformation of the agro-industrial sector from a predominantly subsidy-based model to a more market-oriented configuration of sustainable

development, accompanied by increased sensitivity to external economic conditions.

Phase 3: Institutional stabilization (2018 – 2020). This period is characterized by the recovery of the integral index of sustainable investments to the level of $SII \approx 0.39$. The key factors for stabilization were the equalization of state support ($\mu_{PSE} \approx 0.49$) and favorable medium-term market forecasts ($\mu_{\mathrm{PSO}} \approx 0.64$), which allowed the agro-industrial sector to maintain investment stability even in the context of the pandemic crisis. The coordinated dynamics of economic and environmental components indicate the formation of a phase of relative "stable plateau" characterized by increased adaptive resistance of the system.

Phase 4: Climate volatility and investment resilience (2021–2023). During this period, the model identified a pronounced compensation effect, reflecting the system's adaptive resilience to climate shocks. The droughts of 2021 and 2023 led to a decline in the forecast component ($\mu_{\mathrm{PSO}} \approx 0.47$)), increasing the uncertainty of production expectations. Despite this, the integral index of sustainable investments remained relatively stable at $SII \approx 0.37$ due to the growth of private investment flows and the maintenance of institutional mechanisms, which indicates the formation of investment resilience in the agro-industrial sector in conditions of climate volatility.

A comparative analysis of the dynamics for 2010 and 2023 demonstrates a transition from a predominantly centralized model of investment sustainability based on the dominance of state support to a more polycentric configuration of sustainable development. While in the early stages the dynamics of the index were largely determined by subsidies to the sector, in the final phase, the decline of one factor (climate risks) is offset by the growth of another (private investment). This result confirms the ability of the proposed hybrid fuzzy modular model to identify complex nonlinear relationships between economic, environmental, and investment factors that remain poorly distinguishable within the framework of traditional deterministic analysis methods.

D. Sensitivity Analysis and Model Robustness Verification

Since the weight coefficient ω_j in the AHP method are determined based on expert judgments; the model is subject to epistemic uncertainty. To eliminate the subjective factor and confirm the reliability of the results obtained, a two-stage sensitivity analysis was performed. The purpose of this stage is to test the hypothesis that variations in weights do not lead to critical changes in the ranking of periods (Rank Reversal Problem), which is a key criterion for the quality of decision support systems.

Scenario modeling (Multi-Scenario Strategy):

Four alternative weight distribution scenarios were formed, modeling different investment behavior strategies (Table VII):

- Scenario A (Baseline/AHP): Weights obtained in Section 3.2 ($CR < 0{,}1$). Reflects the current consensus of experts.
- Scenario B (Profit-Maximization): Aggressive economic strategy. The total weight of financial indicators (PSE + FDI) is increased to 0.8.
- Scenario C (ESG-Oriented): Sustainable development strategy. The weight of the environmental factor (AEI) dominates ($\omega = 0.5$), mimicking the strict requirements of "green" funds.
- Scenario D (Equal Weights): Balanced distribution, excluding any preferences ($\omega_j = 0.25$).

TABLE VII. DYNAMICS OF THE INTEGRAL INDEX (SII 2023) DEPENDING ON THE WEIGHT DISTRIBUTION SCENARIO

Scenario	ω_{PSE} (Support)	ω_{FDI} (Investments)	ω_{AEI} (Environment)	ω_{PSO} (Forecast)	SII 2023	Deviation from Base (Δ)
A. Base (AHP)	0.46	0.28	0.09	0.17	0.292	—
B. Profit-Max	0.5	0.30	0.05	0.15	0.301	+3.1
C. ESG-Oriented	0.20	0.10	0.5	0.2	0.373	+27.7
D. Equal Weights	0.25	0.25	0.25	0.25	0.345	+18.1

The simulation results show high structural stability of the model. Despite the fact that in the "green" scenario (C) the absolute value of the index increased (due to the fact that in 2023 the environmental load $\mu_{AEI} = 0.37$ was higher than the decline in investment $\mu_{FDI} = 0.14$), there was no inversion of ranks between periods. In all scenarios, 2023 is identified as a recovery phase compared to the collapse of 2021. This confirms that the model's conclusions are based on fundamental data rather than weight manipulation.

Diagram analysis: As can be seen in Fig. 2, the scenarios formed cover fundamentally different areas of the decision space. The "Profit-Max" scenario (green line) shows a pronounced asymmetry towards economic axes (PSE, FDI), effectively ignoring environmental risks. In contrast, the ESG-Oriented scenario (red line) forms an "elongated" profile in the direction of the AEI axis, mimicking the behavior of a socially responsible investor. The baseline AHP scenario (blue fill) occupies an intermediate position, providing a balanced consideration of all factors. The fact that the final ranking of periods remains unchanged despite such radical geometric differences in profiles (from purely economic to purely environmental) serves as visual proof of the high structural stability of the proposed model.

Elasticity analysis (OAT Approach): The One-at-a-Time method was used to quantitatively assess the impact of each factor. The sensitivity coefficient (Sc) was calculated, showing the percentage change in the output index (SII) when the input weight of the criterion (ω_j)) changes by ±20%:

$$S_c = \frac{\%\Delta SII}{\%\Delta \omega_j}$$

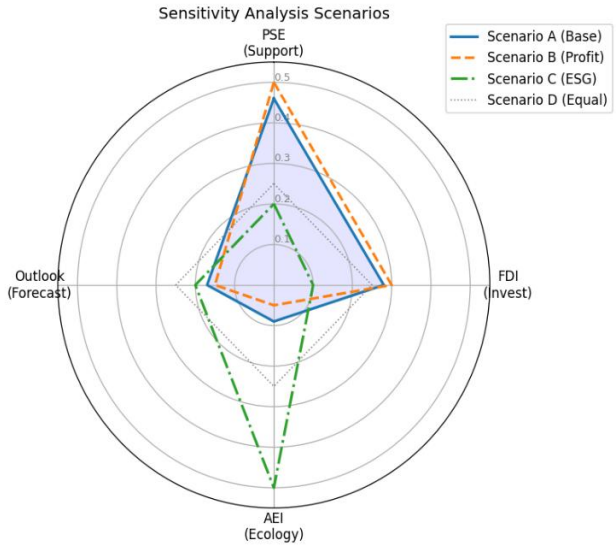


Fig. 2. Visualization of weight distribution profiles for four sensitivity analysis strategies.

The results of the analysis showed:

The greatest elasticity ($S_C \approx 0.42$) was found in the PSE (Government Support) indicator. This mathematically confirms the hypothesis that, in the current configuration of Kazakhstan's agro-industrial complex, government subsidies remain the main driver of investment attractiveness.

A dampening effect ($S_C \approx 0.08$) was observed for the AEI (Ecology) indicator. Low sensitivity to this parameter in the baseline scenario indicates that current environmental issues are not yet a critical barrier for investors, but their weight increases sharply when moving to ESG scenarios.

The developed fuzzy model demonstrates robustness. The integral index responds to parameter changes in a predictable and monotonic manner, without chaotic jumps, which makes the system suitable for use in strategic planning, where the cost of forecasting errors is extremely high.

E. Comparative Analysis with Traditional Methods (Benchmarking)

To assess the added analytical value of the proposed fuzzy multi-criteria approach, a comparison was made with the classical deterministic method of additive weighting (Crisp Simple Additive Weighting, SAW), in which normalized input indicators were aggregated without the use of fuzzyfaction.

Fig. 3 shows the dynamics of the indices calculated by both methods on the same data set for 2010–2023.

Analysis of comparison results: The comparative analysis revealed a fundamental difference in the behavior of the models:

Smoothing Effect: The traditional Crisp method (gray dotted line) demonstrates high signal noise. It reacts with sharp peaks to short-term events: for example, the abnormal jump in 2015 (caused by a one-time inflow of FDI) or the slump in 2012 (drought). Such volatility $\left(\sigma_{crisp} \approx 0.18\right)$ generates many false signals, making it difficult to make long-term decisions;

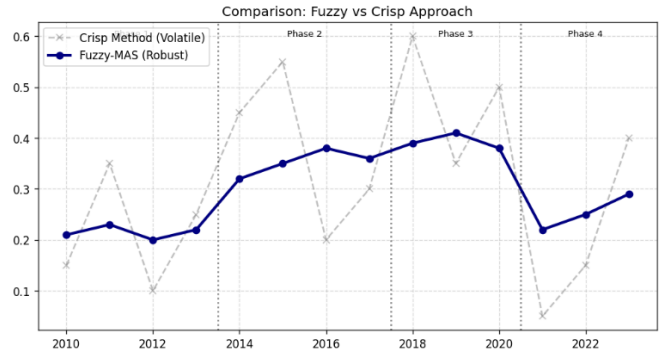


Fig. 3. Comparative analysis of index dynamics: the proposed Fuzzy SII method (smoothed trend) versus the traditional Crisp Index method (high volatility).

The model ignores small changes within the fuzzy set core, responding only to significant structural shifts. This made it possible to identify a stable long-term growth trend for the period 2018 – 2020, free from stochastic noise $\left(\sigma_{fuzzy} \approx 0.09\right)$.

Thus, the proposed hybrid fuzzy multi-criteria approach provides a more stable and interpretable basis for strategic investment analysis compared to deterministic additive models that are sensitive to fluctuations in input indicators in noisy and fragmented data.

IV. Discussion

The results confirm that the proposed hybrid computational framework of agent-oriented fuzzy multi-criteria evaluation provides a stable and interpretable integration of heterogeneous economic, environmental, and investment indicators in the analysis of sustainable investments in the agro-industrial sector. Unlike traditional deterministic aggregation methods, which are highly sensitive to short-term fluctuations in statistical data, the developed model forms a smoothed integral index that allows identifying long-term structural trends in investment sustainability while maintaining transparency of the contribution of individual factors. This property is of fundamental importance for strategic planning in agricultural systems operating in conditions of high climatic and institutional uncertainty.

One of the key analytical results of the study is the identification of structural divergence in sustainability factors for the period 2021–2023. While the growth of foreign direct investment within traditional linear models can be interpreted as an unambiguously positive investment signal, the proposed fuzzy model recorded a relative stabilization of the integral index of sustainable investment at a moderate level. This effect is due to the simultaneous increase in institutional and climate risks, reflected in a decrease in the state support indicator. Thus, the model demonstrates the ability to take into account the asynchronous dynamics of factors, forming a more comprehensive and cautious assessment of investment conditions compared to methods focused exclusively on the volume of investment flows.

A comparative analysis with deterministic aggregation methods showed that the use of membership functions can significantly reduce the sensitivity of the integral index to short-

term climate and price shocks. Unlike classical additive models, the fuzzy approach focuses on fundamental long-term trends, reducing the influence of high-frequency statistical noise. This property is particularly important for the agro-industrial sector, where investment decisions are characterized by long time horizons, and the excessive reaction of analytical models to short-term fluctuations can lead to a distortion of strategic priorities for sustainable development.

The results of the sensitivity analysis confirm the structural robustness of the proposed index. The absence of rank inversions with significant variations in weight coefficients, including scenarios with enhanced consideration of ESG factors, indicates that the identified patterns are due to the fundamental dynamics of the input data rather than specific model parameter settings. The conclusions obtained are consistent with recent studies emphasizing the need to integrate environmental constraints into sustainable development investment models and demonstrating the effectiveness of hybrid fuzzy approaches in analyzing risks and uncertainties [26-29].

The use of agent-oriented modular architecture made it possible to solve the key engineering problem of data interoperability, ensuring reproducible integration of international statistical sources. The modular structure of the system increases the scalability of the solution and makes it possible to adapt it to different national contexts without changing the logic of analytical processing. Thus, the contribution of the study lies not in the development of a new fuzzy operator, but in the formalization of a reproducible integrated analytical process focused on the tasks of sustainable investment analysis in the agro-industrial sector.

Despite the results obtained, the study has a number of limitations. First, the use of aggregated national data may smooth out regional differences in agricultural development. Second, the use of macroeconomic forecasts from international organizations limits the consideration of local production risks. Third, the use of symmetric triangular membership functions simplifies the real nonlinear dependencies of investment sustainability. These limitations determine the directions for further research, including the integration of regional data, the expansion of scenario analysis, and the use of adaptive or neuro-fuzzy membership functions to increase the sensitivity of the model to local structural changes.

V. Conclusion

This study developed, implemented, and empirically validated a hybrid intelligent decision support system for assessing sustainable investments in the agro-industrial sector under conditions of high uncertainty, information asymmetry, and volatility in emerging markets. The results confirm that the integration of agent-oriented modular data processing architecture and fuzzy multi-criteria aggregation methods allows for the formation of sustainable and interpretable investment indicators that adequately reflect structural changes in the institutional, economic, and environmental environment.

The main scientific and practical results of the study can be summarized as follows. The developed three-level data processing architecture (Data Modules – Analytical Modules – Aggregation Module) ensured the semantic, temporal, and metric consistency of statistical indicators coming from heterogeneous international and national sources (OECD, FAO, national statistics). The implemented mechanism for normalizing and synchronizing time series made it possible to form a reproducible analytical pipeline that is resistant to omissions, outliers, and statistical noise.

A comparative analysis with deterministic aggregation methods showed that the use of a fuzzy multi-criteria approach based on Fuzzy SAW provides more stable dynamics of the integral index in relation to short-term climatic and market shocks. Unlike classical additive models, which are sensitive to high-frequency fluctuations in input indicators, the proposed system allows identifying fundamental long-term trends in investment stability. This property is of fundamental importance for the agro-industrial sector, which is characterized by long investment cycles and high climate sensitivity.

Empirical testing of the model using data from the agro-industrial sector of the Republic of Kazakhstan for the period 2010–2023 revealed a qualitative structural shift in the investment environment. In particular, the system diagnosed an institutional imbalance in 2021–2022, when the recovery of market indicators was accompanied by a decline in the integral assessment of sustainable investments due to a reduction in state support. This result confirms the ability of the proposed model to identify hidden regulatory and institutional risks that remain difficult to distinguish using traditional linear methods of investment analysis.

The results of the sensitivity scenario analysis demonstrated the high structural robustness of the model. Variations in the weighting coefficients of the criteria across a wide range, including scenarios with priority consideration of ESG factors, did not lead to an inversion of the ranks of the analyzed periods. This indicates that the estimates obtained are determined by the fundamental dynamics of the input data, rather than subjective parameter settings, which is a critical requirement for strategic investment decision support systems.

The developed system can be considered a practical tool for strategic scenario analysis and support for sustainable investment policy. It can be used by government agencies to assess the impact of changes in the subsidy and regulatory system, as well as by private investors and analysts to make informed investment decisions that are less susceptible to short-term market noise.

The results obtained go beyond the national context. The modular and data-source-agnostic architecture allows the proposed approach to be adapted to the statistical systems of other countries with emerging economies while maintaining a unified logic for integrating economic, environmental, and institutional factors. This opens up opportunities to use the developed framework for cross-country comparative analysis and the formation of supranational systems for monitoring sustainable investments in the agro-industrial sector.

Further research could focus on increasing the spatial detail of the analysis, moving from the country to the regional level, and integrating adaptive neuro-fuzzy methods for calibrating membership functions. Another promising direction is the expansion of the analytical architecture by including machine

learning forecasting modules to assess local climate and production risks in the context of sustainable development.

REFERENCES

[1] OECD. Agricultural Policy Monitoring and Evaluation 2024: Economic and Environmental Roles of Agricultural Policies; OECD Publishing: Paris, France, 2024.

[2] OECD/FAO. OECD-FAO Agricultural Outlook 2023–2032; FAO: Rome, Italy; OECD Publishing: Paris, France, 2023.

[3] M. Lezoche, J. E. Hernandez, M.d.M.E.A. Díaz, H. Panetto and J. Kacprzyk "Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture," Computers in Industry, p.117, 2020.

[4] A. Subeesh, N. Chauhan "Agricultural digital twin for smart farming: A review," Green Technologies and Sustainability, 4(2), 100299, 2026.

[5] R. Lassoued, D.M. Macall, S.J. Smyth, P.W.B. Phillips, H. Hesseln "Expert Insights on the Impacts of, and Potential for, Agricultural Big Data. Sustainability," 13, p.2521, 2021.

[6] P. Martin "Post-Soviet agricultural restructuring: A success story after all?," Comparative Economic Studies, ISSN 1478-3320, Palgrave Macmillan, Basingstoke, vol. 63(4), p.623-647, 2021.

[7] M. Svanidze, L. Götz "Determinants of spatial market efficiency of grain markets in Russia," Food Policy, vol. 89, 101769, 2019.

[8] M.E. Pérez-Pons, R.S. Alonso, O. García, G. Marreiros, J.M. Corchado "Deep Q-Learning and Preference Based Multi-Agent System for Sustainable Agricultural Market," Sensors 21, p.5276, 2021.

[9] A. Michailidis, K. Mattas, "Using Real Options Theory to Irrigation Dam Investment Analysis: An Application of Binomial Option Pricing Model," Water Resources Management: An International Journal, Published for the European Water Resources Association (EWRA), Springer;European Water Resources Association (EWRA), vol. 21(10), p. 1717-1733, October, 2007.

[10] S. Lim 'Risks in Agricultural Supply Chains (University of Chicago Press, 2021)", chap. 2,

[11] P.M. Clay, R. Feeney, J. Sellare 'Technology-driven transformations in agri-food global value chains: The role of incumbent firms from a corporate venture capital perspective," Food Policy, vol. 127, 102684, 2024.

[12] B. Wang, Ch. Wei, L. Shi, X. Cheng, X. Shi "ESG and firm operational efficiency: evidence from Chinese listed companies," Environment, Development and Sustainability: A Multidisciplinary Approach to the Theory and Practice of Sustainable Development, Springer, vol. 27(1), p. 681-714, January, 2025.

[13] D. Yiqun, Y. Fan, X. Lin. 'Environmental, Social, and Governance (ESG) Performance and Firm Value: Evidence from Chinese Manufacturing Firms. Sustainability," vol.15, p.12858, 2023.

[14] B. Wang, M. J. Yang, A Study on the Mechanism of ESG Performance on Corporate Value—Empirical Evidence from A-Share Listed Companies in China. Soft Science, vol.36, p.78-84, 2022.

[15] Zh. Zhai, J. F. Martínez, V. Beltran, N. L. Martínez "Decision support systems for agriculture 4.0: Survey and challenges" Computers and Electronics in Agriculture, vol. 170, 105256, 2020.

[16] L.A. Zadeh "Fuzzy sets," Information and Control, vol 8(3), p.338-353, 1965.

[17] A. Mardani, A. Jusoh, E.K. Zavadskas 'Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014," Expert Systems with Applications, vol. 42(8), p.4126-4148, 2015.

[18] U. Makhazhanova, S. Kerimkhulle, A. Mukhanova, A. Bayegizova, Z. Aitkozha, A. Mukhiyadin, B. Tassuov, A.; Saliyeva, R. Taberkhan, G. Azieva "The Evaluation of Creditworthiness of Trade and Enterprises of Service Using the Method Based on Fuzzy Logic," Appl. Sci. vol.12, 11515, 2022.

[19] T.L. Saaty "Decision Making with the Analytic Hierarchy Process," International Journal of Services Sciences, 1, p.83-98. 2008.

[20] FAO. The State of Food and Agriculture 2023: Revealing the true cost of food to transform agrifood systems; FAO: Rome, Italy, 2023.

[21] M. Mukasheva, A. Mukhiyadin, U. Makhazhanova, and S. Serikbayev "The Behaviour of the Ensemble Learning Model in Analysing Educational Data on COVID-19," IJIET, vol.13(12), p.1868-1878, 2023.

[22] A. Almeida, A. Dale, R. Hay, Y. Everingham, S. Lockie "Environmental, social and governance (ESG) in agriculture: trends and gaps on research," Australasian Journal of Environmental Management. vol.32. p.1-30, 2024.

[23] S. Ibbotson "The Central Asian Economies in the Twenty-First Century: Richard Pomfret," Princeton University Press, New Jersey, 2019. pp. 304.

[24] G. Kou, D. Ergu, Y. Chen, C. Lin, "Pairwise comparison matrix in multiple criteria decision making," Technological and Economic Development of Economy, vol. 22(5), p.738-765, 2016.

[25] S. Kerimkhulle, Z. Dildebayeva, A. Tokhmetov, A. Amirova, J. Tussupov, U. Makhazhanova, A. Adalbek, R. Taberkhan, A. Zakirova, A. Salykbayeva "Fuzzy Logic and Its Application in the Assessment of Information Security Risk of Industrial Internet of Things," Symmetry, vol.15, p.1958. 2023.

[26] A. Nikas, H. Doukas, L. M. López "A group decision making tool for assessing climate policy risks against multiple criteria," Heliyon, vol.4(3), 2018.

[27] S. Zhang, H. Zhang, F. Xie, D. Wu "Climate Change and Sustainable Agriculture: Assessment of Climate Change Impact on Agricultural Resilience," Sustasinability vol.17, p.7376, 2025.

[28] A. Mosavi, P. Ozturk, K.-w. Chau "Flood Prediction Using Machine Learning Models: Literature Review," Water, vol.10, p.1536, 2018.

[29] G.-S. Alberto, F.-S. Juan, O. Waldo. "Predictive ability of machine learning methods for massive crop yield prediction," Spanish Journal of Agricultural Research, vol.12. p.313-328, 2014.

Phonetic Completeness Over Prosodic Diversity: Syllable-Level Synthetic Corpus Construction for Low-Resource Penang Hokkien Speech Synthesis

Yu Liang Lai[1], Yen Min Jasmina Khaw[2], Seng Poh Lim[3], Tien Ping Tan[4]
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, Perak, Malaysia[1, 2, 3]
School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia[4]

***Abstract*—This study presents the first Text-to-Speech (TTS) model for Penang Hokkien, a low-resource tonal dialect at risk of extinction. To address phonological sparsity in the collected speech corpus, we propose a two-stage fine-tuning approach that emphasizes comprehensive phonetic coverage through syllable-level synthetic augmentation while subsequently refining prosodic naturalness using real speech recordings. By supplementing a limited 45-minute real speech corpus with a 2-hour syllable-level concatenative synthetic corpus, the full dialectal inventory of approximately 2,000 unique syllable-tone combinations was encompassed. Experimental results suggest that improving syllable-tone coverage contributes substantially to intelligibility and tonal accuracy in this low-resource tonal setting. Technical optimizations, including a 600-ms cross-fading technique to mitigate boundary artifacts and numerical tone markers to reduce token sparsity, further improved model stability and synthesis quality. The final model achieved a Mean Opinion Score (MOS) of 3.92.**

***Keywords*—Low-resource language; speech synthesis; text-to-speech; data augmentation techniques**

I. Introduction

Penang Hokkien is a local variant of Hokkien (Chinese dialect) spoken in Penang, Malaysia. This dialect has evolved uniquely over time, incorporating loanwords from Malay, English, Cantonese, and Teochew, reflecting Penang's rich multicultural heritage [1]. In recent years, the usage of Penang Hokkien has declined, particularly among the younger generation, as Mandarin increasingly dominates communication within the Malaysian Chinese community. This shift places Penang Hokkien at risk of extinction if efforts are not made to preserve it for future generations [2].

One viable approach to safeguarding this dialect is through comprehensive documentation, in terms of text and speech, by building a TTS system. However, developing TTS systems for low-resource languages presents significant challenges, primarily due to the lack of high-quality paired text and speech data. As a predominantly spoken, unwritten dialect, Penang Hokkien lacks a standardized orthography and a comprehensive speech corpus. Most state-of-the-art neural TTS models require extensive high-quality datasets that are infeasible for such low-resource scenarios. Furthermore, even when limited real speech is recorded, it often suffers from phonological sparsity, failing to cover the exhaustive inventory of possible syllable-tone combinations, which results in poor performance on out-of-vocabulary (OOV) terms.

To mitigate the resource constraints, this study implements a two-stage fine-tuning approach centered on phonetic completeness leveraging the SpeechT5_TTS model architecture [3]. By establishing a dense phonetic foundation through a hybrid speech corpus and optimizing tonal representation, this work proposes a potentially scalable framework and serves as an initial case study for the digital preservation of unwritten tonal dialects.

The major contributions of this study are as follows:

1) We empirically investigate the relationship between phonetic coverage and prosodic variation in low-resource tonal TTS. Experimental results suggest that expanding syllable-tone coverage through syllable-level concatenative augmentation contributes more consistently to intelligibility and tonal accuracy than prosodic perturbation alone, particularly for out-of-vocabulary (OOV) syllables.

2) We propose and validate the use of numerical tone markers as a superior alternative to traditional diacritics for low-resource TTS. We demonstrate that this representation reduces token sparsity by approximately 35% and accelerates model convergence by 18%, providing a more stable embedding space for rare syllable units.

3) We implement a two-stage fine-tuning approach using a hybrid speech corpus with an identified optimal 2:1 synthetic-to-real data ratio. Additionally, a 600-ms cross-fading technique is introduced to successfully mitigate the boundary artifacts inherent in concatenative synthesis, balancing phonetic robustness with natural human prosody.

4) To the best of our knowledge, this study establishes the first functional TTS framework for the Penang Hokkien dialect. This provides a vital technological foundation for the digital preservation of this endangered unwritten language.

II. Literature Review

A. Low-Resource TTS Dataset Acquisition

Recent advancements in neural speech technologies and the challenges of low-resource scenarios have prompted significant changes in TTS dataset construction. Traditional approaches relying on extensive audio recordings are often impractical for low-resource languages. To address this, researchers are increasingly repurposing existing data to create effective TTS datasets.

For example, LibriTTS [4] and CML-TTS [5] were adapted from ASR datasets, using carefully designed pipelines for text and audio processing to ensure compatibility with TTS systems. These pipelines retained critical features of the original ASR datasets, such as speaker diversity, text-audio alignment, and high-quality recordings, while resolving common issues like lengthy utterances, transcription inconsistencies, and noisy audios. In contrast, projects like the CMU Wilderness Multilingual Speech Dataset [6], BibleTTS [7], and Sanskrit TTS [8] have utilised publicly available Bible readings as a foundation to construct speech corpus. Through meticulous data preprocessing, these initiatives have successfully transformed crawled data into reliable resources for speech synthesis.

Researchers have also explored the potential of other found audio sources. [9] analyzed acoustic and prosodic features of various audio recordings, identifying that broadcast news is particularly suitable for TTS tasks. Similarly, a small TTS corpus for Moroccan dialect was created using storytelling audios from YouTube, which were denoised, segmented, transcribed with ASR system, and manually normalized to ensure accuracy [10]. Furthermore, unsupervised data selection approaches have shown promise in low-resource settings. In [11], researchers used broadcast data to create a TTS dataset, achieving a high Mean Opinion Score (MOS) of 4.4 for intelligibility with only one hour of data, demonstrating the potential of combining broadcast data with selection methodologies to address resource constraints effectively.

B. Low-Resource Speech Synthesis

Various approaches were employed to address data scarcity in low-resource languages. A major direction seeks to reduce reliance on paired text-speech corpora by leveraging representation learning. These approaches exploited large pre-trained models to extract intermediate representations that capture phonetic or acoustic information without requiring explicit labels. For example, language models such as Bidirectional Encoder Representations from Transformers (BERT) [12][13] and self-supervised speech models including Vector Quantized Variational Autoencoder (VQ-VAE) [14][15] and Hidden-Unit BERT (HuBERT) [16] have been employed to learn compact acoustic-symbolic units directly from raw speech. These representations approximate phoneme sequences and serve as intermediate units for training TTS models without extensive labelled data.

Beyond representation learning, unsupervised training has also been explored to bypass the need for parallel datasets altogether. For example, [17] proposed training two separate modules, one for alignment and one for synthesis on non-parallel data, enabling TTS learning without explicit text-speech pairs. Another influential direction exploited the duality between ASR and TTS through closed-loop learning. The "speech chain" framework [18], for instance, co-trains ASR and TTS by allowing TTS-generated audio to be transcribed by ASR, with the transcriptions fed back as supervisory signals to TTS. Similarly, ASR models have been used to produce pseudo-labels for unlabelled speech [19].

Complementary to unsupervised approaches, data augmentation and cross-lingual adaptation form another practical toolkit for low-resource scenarios. Conventional augmentation such as speech perturbation, pitch shifting, and background noise addition [20][21] are widely used to artificially expand existing datasets. While these approaches play a vital role, transfer learning remains the standard approach for low-resource TTS, either used solely or in combination with other approaches mentioned before. Its effectiveness was demonstrated through pretraining on a single high-resource language [22][23]. In more recent works, its capabilities have been extended to include multilingual and multi-speaker modelling for enhanced adaptability [24]. Phonetic similarity in relation to language closeness has also emerged as a key factor in transfer learning. In [25], a hierarchical transfer learning approach revealed that phonologically close languages significantly benefit from multilingual transfer learning, highlighting the importance of nuanced approaches in addressing the challenges of TTS in low-resource scenarios.

III. Methodology

To effectively address the data scarcity and phonological sparsity inherent in the Penang Hokkien dialect, this study proposes a comprehensive methodology that bridges the gap between synthetic data engineering and progressive neural model training. The overall architecture is divided into two primary phases: (a) hybrid speech corpus construction and (b) two-stage text-to-speech (TTS) model training. The first phase utilizes a dual-pipeline approach to acquire raw data, subsequently undergoing acoustic and linguistic processing, including 600-ms linear cross-fading and numerical tone representation, to generate a phonetically-rich synthetic speech corpus and a prosodically-authentic real speech corpus. In the second phase, these corpora are leveraged through a sequential fine-tuning strategy. The process begins with a pretrained SpeechT5_TTS model ($M_{pretrained}$) [3] that undergoes intermediate fine-tuning on synthetic data to establish a phonetic foundation by learning basic tonal contours across the comprehensive syllable inventory ($M_{intermediate}$), followed by final fine-tuning on real speech to achieve prosodic refinement and human-like naturalness (M_{final}) (Fig. 1).

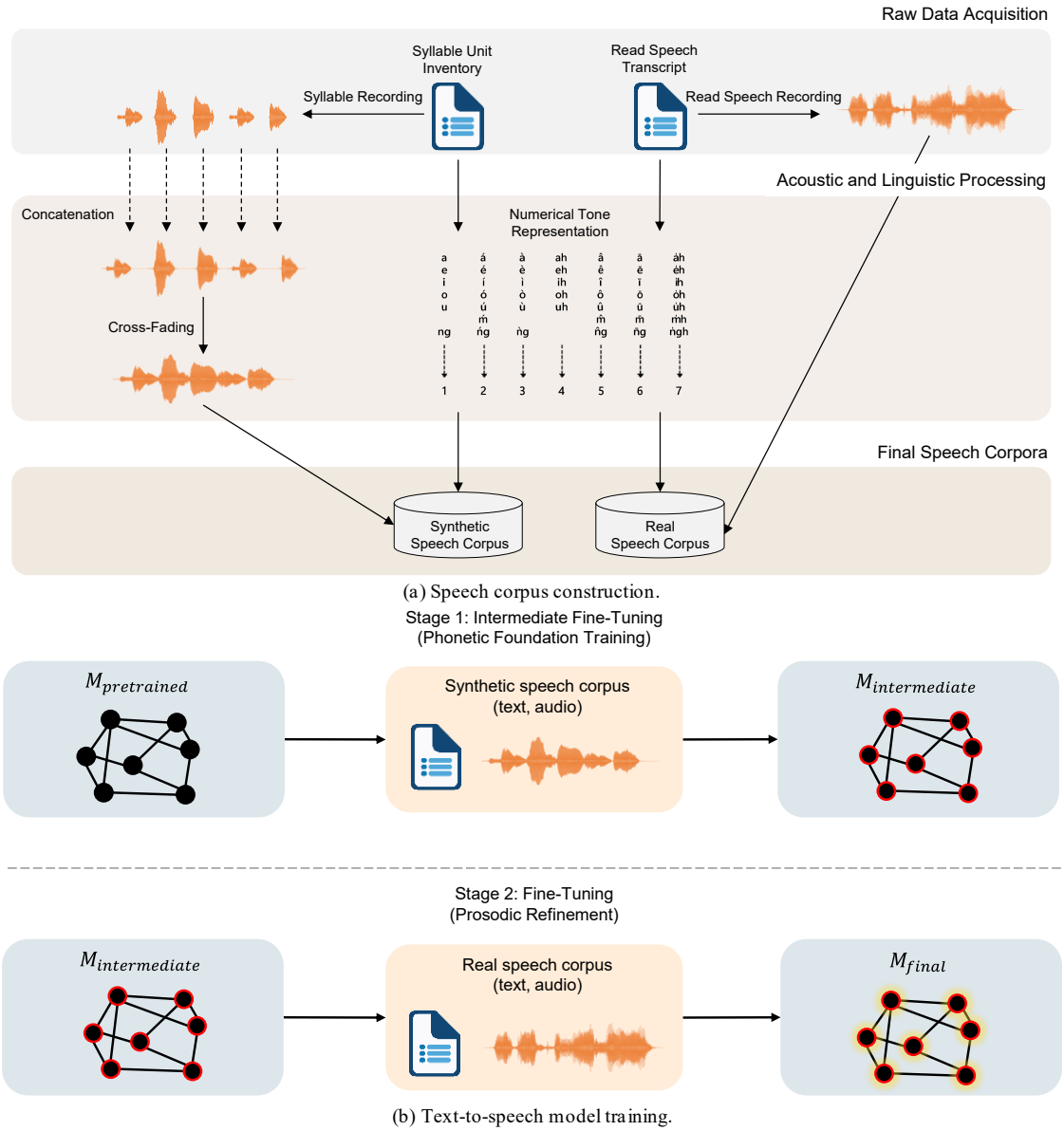


Fig. 1. The Penang Hokkien TTS framework is divided into two distinct components: (a) The speech corpus construction phase, which establishes a hybrid dataset by integrating isolated syllable units and continuous read speech via numerical tone mapping and acoustic smoothing; and (b) The text-to-speech model training phase, illustrating the progressive two-stage fine-tuning strategy used to adapt a pretrained model into a phonetically-complete and prosodically-refined final TTS system.

A. Real Speech Corpus: The Acoustic Foundation

The real speech corpus serves as the acoustic foundation for the TTS model. The transcript was sourced from documented conversations and community-driven social media platforms [26][27] to capture authentic linguistic variations. Inconsistencies in non-standard orthography were resolved using the Soundex$_{PH}$ algorithm [28] to ensure phonetic consistency. A native female speaker recorded the transcript in controlled conditions using high-fidelity equipment at 44.1 kHz to capture a broader frequency range, preserving phonetic details like the velar stop *'ak'*, which might be lost at lower sampling rates. Audios were downsampled to 16 kHz to align with the SpeechT5 architecture after undergoing silence trimming and volume normalization. After these procedures, approximately 45 minutes of high-quality Penang Hokkien real speech corpus was obtained (Table I).

TABLE I. DISTRIBUTION OF REAL SPEECH CORPUS

Statistics	Value
Number of utterances	902
Total syllable count	10507
Unique syllable count	1284
Total duration	44 min 12 sec
Minimum length of utterance	1.04 sec
Maximum length of utterance	7.22 sec
Average length of utterance	1.94 sec

B. Synthetic Speech Corpus: Bridging the Phonological Gap

To bridge the phonological gap (around 700 missing syllables in the 45-min real speech corpus), a 4-hour synthetic speech corpus was constructed (Table II). A synthetic transcript was generated through the random concatenation of romanized syllables to ensure comprehensive coverage of the approximately 2,000 unique syllable-tone combinations in the Penang Hokkien phonetic inventory. Individual syllable units were recorded in isolation by the same native speaker to maintain acoustic consistency. Syllables were utilized as the primary synthesis unit rather than traditional speech unit diphones to better reflect the tonal characteristics of Penang Hokkien while reducing computational complexity [29].

The necessity of the synthetic corpus is highlighted by the phoneme distribution analysis shown in Fig. 2 and 3. The real speech corpus exhibits a heavily skewed distribution where frequent phonemes dominate, leaving a long tail of underrepresented units. Conversely, the syllable-level concatenative approach yields a deliberately balanced distribution, ensuring that the model is exposed to the comprehensive phonetic inventory of Penang Hokkien.

TABLE II. DISTRIBUTION OF THE SYNTHETIC SPEECH CORPUS

Statistics	Value
Number of utterances	5779
Total syllable count	53481
Unique syllable count	2045
Total duration	4 hr 0 min 2 sec
Minimum length of utterance	1.65 sec
Maximum length of utterance	7.08 sec
Average length of utterance	2.24 sec

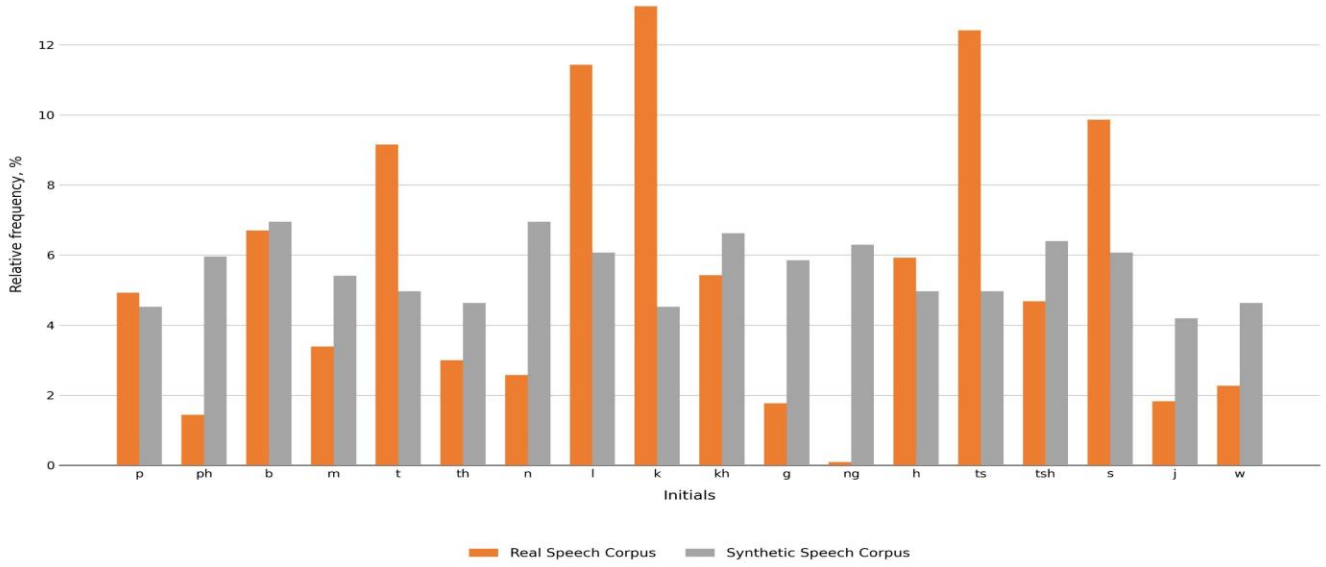


Fig. 2. Initials phoneme distribution.

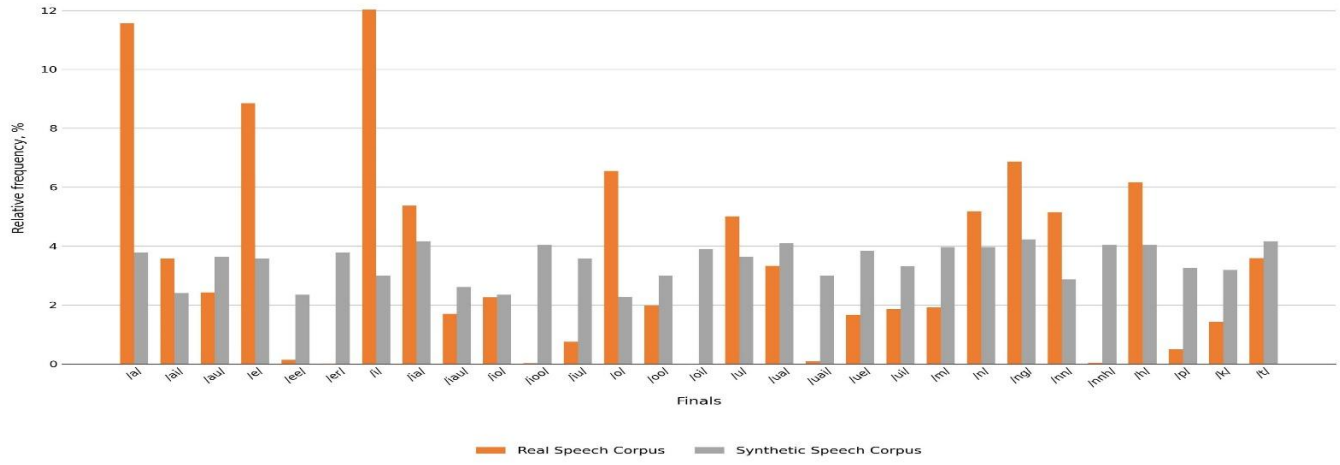


Fig. 3. Finals phoneme distribution.

C. Cross-Fading: Concatenative Artifact Mitigation

Concatenated speech may introduce artifacts at the boundaries between syllables, affecting the overall naturalness and fluidity of the synthesized output. These artifacts arise due to acoustic discontinuities between concatenated units, leading to unnatural transitions. To mitigate this issue, a cross-fading technique [30] was applied, which smooths transitions between syllables by gradually blending overlapping segments. This technique helped reduce abrupt changes and enhance the cohesion and flow of the synthesized speech.

The cross-faded audio signal $S(t)$ is defined as:

$$S(t) = \big(1 - \alpha(t)\big) \cdot A_n(t) + \alpha(t) \cdot A_{n+1}(t) \qquad (1)$$

For this study, the duration of each individual syllable unit signal $A_n(t)$ was fixed at 1.0 s, while the overlap duration T was set to 600 ms. The concatenation was achieved by applying a fade-out factor $(1 - \alpha(t))$ to the preceding syllable segment $A_n(t)$ and a fade-in factor $\alpha(t)$ to the succeeding syllable segment $A_{n+1}(t)$. These factors vary linearly with time (t) across the overlap duration T. Specifically:

- $\alpha(t) = \frac{t}{T}$ increases from 0 to 1 as t progresses through the overlap region, gradually bringing in the succeeding syllable segment.
- $1 - \alpha(t)$ decreases from 1 to 0 over the same time, fading out the preceding syllable segment.

The resulting cross-faded audio signal $S(t)$ smoothly transitioned between syllable units, ensuring natural-sounding audio continuity.

D. Numerical Tone Representation: Improve Tonal Learning

Penang Hokkien is a tonal language where variations in tone significantly alter word meaning. The model must correctly encode and distinguish these tonal variations for effective TTS training. The original Penang Hokkien read speech transcript uses diacritic characters (e.g., *"à"*, *"á"*, *"â"*) to indicate tones, but an alternative numerical tone representation (e.g., *"à"* → *"a3"*) was proposed in this study.

Using numerical tone markers in the input text instead of the original diacritic characters provides several advantages for low-resource TTS. First, the Penang Hokkien real speech corpus is limited to only 45 minutes, and individual diacritic characters may appear infrequently, making it difficult for the model to learn meaningful embeddings for these rare tokens due to limited exposure. In contrast, using numbers to represent tones consolidates tonal information into a smaller, shared set of tokens (e.g., *"à"*, *"è"*, and *"ì"* all map to the tone marker *"3"*). This significantly reduces token sparsity, allowing the model to learn tonal variations more effectively. Additionally, direct diacritic encoding would require 34 unique diacritic variations in the tokenizer, leading to a larger vocabulary size and increased embedding complexity. In comparison, numerical tone markers only require 7 additional tokens (1-7), minimizing vocabulary size and computational overhead. Since each token corresponds to a row in the model's embedding matrix, reducing the number of unique tokens simplifies the embedding space and enhances generalization, particularly in low-resource scenarios. Fig. 4 illustrates the mapping between diacritic tones and their corresponding numerical representations.

Diacritic tone representation :	a e i o u ng	á é í ó ú m̀ ńg	à è ì ò ù ǹg	ah eh ih oh uh	â ê î ô û m̂ n̂g	ā ē ī ō ū m̄ n̄g	a̍h e̍h i̍h o̍h u̍h m̍h n̍gh
Numerical tone representation :	1	2	3	4	5	6	7

Fig. 4. Diacritic tones in Penang Hokkien and their corresponding numerical tone representations.

E. Two-Stage Fine-Tuning: Phonetic Training and Prosodic Refinement

The Penang Hokkien TTS model was trained using a progressive two-stage fine-tuning approach designed to balance broad phonetic robustness with authentic prosodic naturalness. The model was initialized with a multitask-pretrained SpeechT5_TTS checkpoint trained on the high-resource English LibriTTS corpus [3], enabling cross-lingual knowledge transfer to the low-resource Penang Hokkien dialect.

The primary objective of the first stage was to establish a dense phonetic foundation and resolve the model's exposure to the comprehensive inventory of approximately 2,000 unique syllable-tone combinations. Based on our ablation study regarding optimal augmentation volume, the model underwent intermediate fine-tuning on 2 hours of the synthetic speech corpus. This established a 2:1 synthetic-to-real data ratio, which was identified as the inflection point for maximum performance. Before training, the tokenizer was extended to include numerical tone markers (1–7) to prevent the loss of critical tonal information. To accommodate these new tokens, the embedding layer was resized. During this stage, all layers were unfrozen to allow the model to fully integrate the newly established tonal relationships and adapt its hidden representation space to the modified input.

The second stage focused on refining acoustic modeling and speech feature generation while preserving the phonetic structures learned in Stage 1. The model was fine-tuned on the 45-minute real speech corpus to supply the natural prosody and authentic accentual nuances inherent in human recordings. To retain the learned phonetic structures, the token embedding layers, lower transformer layers, and self-attention mechanisms were frozen. Fine-tuning was restricted to the acoustic modeling layers (encoder and decoder) and the decoder's pre-net and post-net layers to optimize spectral feature prediction.

To convert the predicted mel-spectrograms into high-fidelity speech waveforms, we utilized a pretrained HiFi-GAN vocoder [3]. Unlike the acoustic model, the vocoder was not fine-tuned from scratch, as its architecture is designed to generalize fundamental speech characteristics, such as periodicity and loudness, across different languages and speakers [31].

Table III shows the hyperparameter settings used in model training.

TABLE III. MODEL TRAINING HYPERPARAMETERS

Hyperparameter	Value
learning_rate	1e-4
train_batch_size	4
eval_batch_size	2
seed	42
gradient_accumulation_steps	8
total_train_batch_size	32
optimizer	ADAMW_TORCH with β_1, β_2=(0.9,0.999) and ε =1e-8
lr_scheduler_type	linear
lr_scheduler_warmup_steps	500
training_steps	2000
mixed_precision_training	Native AMP

IV. EVALUATION

A. Objective Evaluation

The objective evaluation of synthesized Penang Hokkien speech employs on Mel-Cepstral Distortion (MCD) [32], a widely used spectral distance metric that quantifies the similarity between synthesized and groundtruth speech in terms of acoustic features. MCD measures the distance between the mel-frequency cepstral coefficients (MFCCs) extracted from both speech signals, providing an objective assessment of spectral fidelity. A lower MCD score indicates a smaller spectral difference, implying higher fidelity and better synthesis quality.

B. Subjective Evaluation

For subjective evaluation, Mean Opinion Score (MOS) [33] was conducted to assess the quality of the synthesized Penang Hokkien speech. MOS is a widely accepted perceptual metric that quantifies speech quality on a 5-point scale. To eliminate potential biases, the utterances were presented in a randomized order. A total of 10 native Penang Hokkien speakers participated in the evaluation. A structured questionnaire was utilized to evaluate the synthesized speech across three fundamental dimensions for tonal language synthesis [34]:

- Intelligibility: Measures how well a listener can understand the synthesized speech, including pronunciation accuracy, phoneme articulation, and word clarity.
- Naturalness: Evaluate how human-like the synthesized speech sounds, considering smoothness, voice quality, and absence of robotic artifacts.
- *Tone Accuracy* – Assesses how well the synthesized speech captures intonation, pitch variation, and stress patterns, which are essential in Penang Hokkien as a tonal language (Table IV).

TABLE IV. MEAN OPINION SCORE (MOS) EVALUATION SCALE

Score	Intelligibility	Naturalness	Tone Accuracy
1	No meaning understood	Unnatural, highly robotic	Not accurate, very big tonal shift
2	Effort required to understand	Inadequately natural	Inadequately accurate
3	Moderate effort required	Adequately natural	Adequately accurate
4	No major effort, mostly clear	Near natural	Near accurate, slight tonal shifts
5	No effort required, perfectly clear	Natural human speech	Very accurate, correct tone

V. RESULTS AND DISCUSSION

The primary contribution of this study is the empirical validation that, for tonal dialects like Penang Hokkien, addressing the "phonological gap" through a comprehensive phonetic inventory is more vital for synthesis quality than simply increasing the amount of training data through acoustic perturbations. While common data augmentation techniques like pitch-shifting [35] improve model robustness by simulating prosodic diversity, they do not introduce new phonetic information.

We conducted a controlled comparison between two distinct augmentation techniques, each providing 2 hours of synthetic speech data for intermediate fine-tuning before fine-tuning on the 45-min real speech data, with numerical input as tone representation.

1) Syllable-level concatenation: This augmentation technique constructed synthetic speech to cover the comprehensive set of approximately 2,000 Penang Hokkien syllables, ensuring every tonal and phonetic unit was represented during training.

2) Pitch-shifting: Augmentation was performed on the original single-speaker real speech corpus using the SoX tool [36] to simulate prosodic variation. Pitch was shifted within a narrow range of -0.5 to +0.5 semitones (step size: 0.5), and speed was perturbed by ± 10% (ratios: 0.9 and 1.1). This restricted range was selected to maintain speaker similarity, as larger perturbations are known to degrade perceived voice identity and naturalness [37][38]. While this augmentation technique provided an equivalent 2 hours of synthetic speech data, it remained restricted to the 1,300 syllables present in the original real speech corpus (Table V).

TABLE V. SYLLABLE-LEVEL CONCATENATION VS. PITCH-SHIFTING (PHONETIC COVERAGE VS. PROSODIC VARIATION)

Augmentation Technique	Unique Syllable Count	MOS ↑				MCD ↓
		Intelligibility	Naturalness	Tone Accuracy	Overall	
2 hr Syllable-Level Concatenation (Proposed)	2045	**4.12 ± 0.43**	3.95 ± 0.24	**3.69 ± 0.52**	**3.92 ± 0.40**	**20.07 ± 2.09**
2 hr Pitch-Shifting	1284	3.67 ± 0.26	**4.09 ± 0.17**	3.34 ± 0.51	3.70 ± 0.31	20.54 ± 1.86

The results reveal that the proposed syllable-level concatenation (MOS 3.92) consistently outperformed pitch-shifting (MOS 3.70). Pitch-shifting provides some benefits by introducing prosodic variation, which has been shown in prior work to improve model robustness [37]. However, this technique remains limited because it did not improve phonetic coverage, many syllables remain unseen during training. In contrast, the proposed syllable-level concatenation technique directly addressed phonetic coverage limitations. By systematically constructing synthetic speech to cover the comprehensive set of ~2,000 Penang Hokkien syllables, the model is exposed to every tonal and phonetic unit before final fine-tuning on a real speech corpus. This broader coverage is particularly critical for tonal languages, where accurate tone realization is tightly coupled with intelligibility [39].

Although pitch-shifting yields a slight improvement in perceived naturalness, native speaker evaluators consistently preferred the model intermediately fine-tuned on syllable-level concatenative synthetic speech for its clarity and tonal correctness. The higher ratings in intelligibility and tone accuracy indicate that in a low-resource tonal language setting, comprehensive phonetic coverage contributes more directly to synthesis quality than increased prosodic variation alone.

The superiority of the syllable-level concatenation technique is most evident when synthesizing Out-of-Vocabulary (OOV) syllables, such as *"gēr"*, *"lēeng"*, and *"ngāng"* in the sentence *"tsin gēr-lam khuànn lú tī tsia lēeng-ngāng-lēeng-ngang. (It's frustrating seeing you lengang-lengang here—just doing nothing.)"*. The effectiveness of this technique is visually substantiated through mel spectrogram analysis, which represents the spectral properties and energy distribution of synthesized speech. In Fig. 5, we compare the acoustic outputs for the OOV sentence across different training configurations.

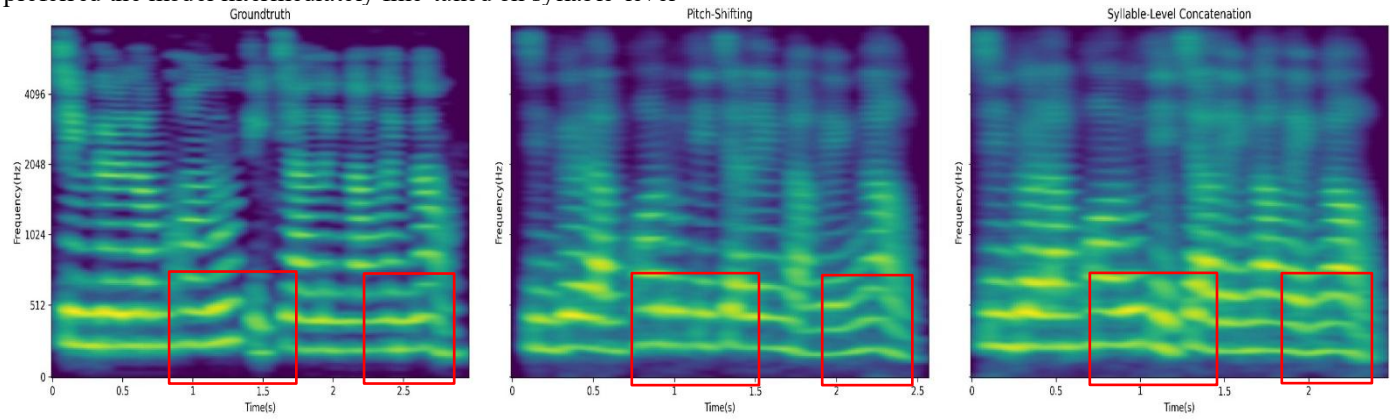


Fig. 5. Mel spectrogram analysis of OOV sentence.

Mel spectrograms represent the spectral properties of speech, where clear and well-defined harmonic structures (the bright horizontal bands) correlate with higher fidelity and intelligibility. The groundtruth spectrogram displays sharp, distinct harmonics, representing clear, natural speech. These features are indicative of high-fidelity natural speech with stable pitch contours. The output from the pitch-shifting model is noticeably blurrier, indicating weak harmonic definition and degraded spectral fidelity. While pitch-shifting introduces prosodic variety, it fails to equip the model with the necessary phonetic information to accurately reconstruct the spectral envelopes of OOV syllables. In contrast, the model trained with syllable-level augmentation produces significantly sharper spectrograms that closely resemble the groundtruth reference speech. The harmonic bands are more pronounced and stable, demonstrating that intermediate fine-tuning on a comprehensive phonetic inventory allows the model to learn correct phonetic and tonal structures for OOV syllables with high accuracy.

While the improved performance is consistent with the hypothesis that broader phonetic coverage benefits low-resource tonal dialect TTS, the observed gains are likely influenced by multiple interacting factors. First, the two-stage fine-tuning strategy may introduce curriculum learning effects, where intermediate training on synthetic speech establishes stable phonetic representations before refinement on natural speech. Second, the synthetic corpus produces denser token distributions by repeatedly exposing the model to shared syllable-tone patterns, potentially improving embedding stability for rare units. Third, intermediate pretraining on synthetic speech may also provide a regularization effect by exposing the model to a wider range of phonetic combinations prior to adaptation on the limited real speech corpus. Therefore, while the experiments suggest that phonetic completeness plays an important role, the individual contributions of these factors cannot yet be fully disentangled within the current study design.

VI. Ablation Studies

A. Cross-Fading on Concatenative Artifact Mitigation

During the preparation of the synthetic speech corpus, individual syllable units were recorded and concatenated to form short utterances. However, naive concatenation often results in audible discontinuities at unit boundaries, such as clicks, spectral jumps, and unnatural transitions, which can negatively impact downstream TTS training. To ensure the synthetic corpus provides a clean phonetic foundation, we evaluated the effectiveness of cross-fading technique in smoothing the acoustic transitions within the syllable-level concatenative synthetic speech.

We compared direct concatenation (0 ms) against a 600-ms cross-fade across various phonotactic contexts. To objectively quantify join quality, we utilized boundary-level Mel-Cepstral Distortion (MCD) [40] to measure spectral mismatch at the transition points (as primary source of perceptual artifacts).

TABLE VI. ACOUSTIC AND PERCEPTUAL EFFECTS OF CROSS-FADING

Condition	MOS (Naturalness) ↑	Boundary MCD (dB)↓
No cross-fading	3.07 ± 0.44	6.41 ± 0.92
600ms cross-fading	3.68 ± 0.38	3.74 ± 0.51

As shown in Table VI, the 600-ms cross-fading reduced boundary MCD from 6.41 dB to 3.74 dB, an improvement of approximately 41% compared to direct concatenation (no cross-fading). High MCD values in the no-cross-fading condition reflect abrupt spectral jumps, a well-documented artifact in concatenative synthesis. Applying a 600-ms cross-fading mitigated these jumps by gradually interpolating between spectral envelopes, yielding smoother formant trajectories and suppressing spurious high-frequency energy at the join. This smoothing also supports more natural F0 continuity across syllables. This reduction in spectral discontinuity translated to significant perceptual gains, with MOS ratings increasing from 3.07 (unnatural) to 3.68 (moderately natural). Evaluators reported a marked reduction in audible clicks and smoother transitions. These results confirm that cross-fading is an essential preprocessing step to ensure the model learns acoustic transitions consistent with natural coarticulation. Consequently, the 600-ms configuration was adopted for the final synthetic corpus construction.

B. Numerical vs. Diacritic Tone Representation Markers

In tonal dialects like Penang Hokkien, the tone markers used to represent pitch variations in the input text is a critical design factor that directly influences the model's ability to learn phonetic patterns. While diacritics are linguistically traditional, they often introduce high token sparsity and vocabulary complexity in low-resource scenarios, where data is insufficient to train rare character embeddings. We conducted this study to determine if numerical markers could simplify the embedding space and accelerate model's convergence without compromising tonal information.

The input text with diacritic tone markers (e.g. "à") and numerical tone markers (e.g. "a3") was compared across four dimensions:

1) Vocabulary coverage: The use of diacritic tone markers resulted in a larger and sparser token vocabulary, as each of the 34 unique diacritic variant was treated as a distinct token. Within the 45-minute real Penang Hokkien speech corpus, many of these diacritic variants appear only a few times, producing a long tail of low-frequency tokens. Numerical tone markers collapsed all these diacritic variants into a small, shared set of tone markers (1-7), significantly reducing tokenizer vocabulary by ≈35%.

2) Model convergence speed: Reusing a small set of numerical tone markers created a denser token distribution, allowing the model to learn tone embeddings earlier and with greater stability. Diacritics, conversely, resulted in many low-frequency tokens that produced noisy gradients and slowed down the optimization of embeddings. Since numerical tone markers appeared across many syllables, the model was exposed to more syllable-tone combinations for each tone class, enabling more consistent mappings between tone markers and intonation. Quantitatively, the numerical model consistently reached a stable validation loss plateau (0.5936) at Step 950 (Epoch≈39), compared to (0.5939) at Step 1150 (Epoch≈48) for the diacritis baseline, representing a faster convergence speed of ≈18%.

3) Synthesis tonal accuracy: Tonal accuracy was assessed by analyzing synthesis output. Tone substitution error was identified whenever the synthesized tone class did not match the expected tone. The numerical model achieved a significant reduction in error rates (6.5%) compared to diacritic model (18.2%).

4) Perceptual speech quality (XAB Listening Test) [41]: A preference test involving native speakers revealed a clear preference for the numerical model. 72.45% of responses favored the numerical-tone outputs due to more stable tonal contours and higher intelligibility, while 0% favored the diacritic model. The remaining 27.55% of responses indicated no preference, representing instances where perceptual differences were negligible.

These results are summarized in Table VII. These findings confirm that numerical tone markers are superior to diacritic tone markers for Penang Hokkien TTS model training, effectively improving both computational efficiency and human perceptual preference.

TABLE VII. NUMERICAL VS. DIACRITIC TONE MARKERS

Evaluation Dimension	Numerical Tone Markers	Diacritic Tone Markers	Outcome
Vocabulary coverage	Reduced by ~35% consolidated shared tokens	Larger, sparse token set due to infrequent diacritic forms	Numerical better (less sparsity)
Model convergence speed	~18% fewer steps to converge	Slower, converged in more training steps	Numerical faster
Tone mispronunciation rate	6.5% error rate	18.2% error rate	Numerical more accurate
Perceptual preference (XAB test)	72.45% preference (27.55% no preference)	0% preference	Numerical strongly preferred

C. Optimal Volume of Syllable-Level Concatenative Synthetic Speech Data

A critical challenge in low-resource TTS is identifying the "inflection point" where synthetic data provides maximum phonetic stability without introducing robotic artifacts or overfitting to mechanical prosody of concatenated units. We investigated this by fine-tuning four configurations with varying durations of the synthetic speech corpus (0-4 hours) followed by a constant 45-minute real speech corpus.

Among all evaluated configurations, Model III (2hr synthetic + 45min real) achieved the best overall performance, recording the highest MOS (3.92) and lowest MCD (20.07), indicating the closest spectral match to groundtruth reference speech. These results suggest that supplementing limited real speech data with syllable-level concatenative synthetic speech data helps the TTS model to learn better pronunciation patterns and tonal variations, yielding clearer and more accurate speech synthesis while maintaining naturalness (Tables VIII and IX).

TABLE VIII. TTS TRAINING DATA CONFIGURATIONS

Model	Synthetic Speech Data	Real Speech Data	Number of utterances	Total Syllable Count	Unique Syllable Count	Input Tone Representation
Model I	-	45 min	902	10507	1284	Numerical
Model II	1 hr	45 min	2110	13472	2045	Numerical
Model III	2 hr	45 min	3302	26502	2045	Numerical
Model IV	4 hr	45 min	5779	53481	2045	Numerical

TABLE IX. OBJECTIVE AND SUBJECTIVE EVALUATION OF SYNTHESIZED SPEECH

Model	MOS ↑				MCD ↓
	Intelligibility	Naturalness	Tone Accuracy	Overall	
Model I	3.71 ± 0.67	3.62 ± 0.42	3.38 ± 0.63	3.57 ± 0.57	21.04 ± 2.45
Model II	3.89 ± 0.39	3.78 ± 0.37	3.55 ± 0.44	3.74 ± 0.40	20.21 ± 1.97
Model III	**4.12 ± 0.43**	**3.95 ± 0.24**	3.69 ± 0.52	**3.92 ± 0.40**	**20.07 ± 2.09**
Model IV	3.92 ± 0.31	3.53 ± 0.33	**3.73 ± 0.45**	3.73 ± 0.36	20.14 ± 2.11

However, increasing the amount of synthetic speech data beyond an approximately 2:1 ratio did not yield meaningful improvements. Model IV (4hr synthetic + 45min real) showed only a marginal increase in tone accuracy (MOS 3.73 vs 3.69), but experienced slight declines in intelligibility and naturalness, and a higher MCD score of 20.14 ± 2.11. These results indicate a performance plateau, with signs of degradation beyond the optimal augmentation point. Several factors likely contributed to this degradation. First, syllable-level concatenative synthetic speech may introduce unnatural prosody and boundary artifacts. With more synthetic speech data, the model might overfit to these unnatural patterns, negatively affecting speech smoothness. Second, the phonetic diversity of synthetic speech data tends to diminish after a certain threshold; beyond two hours, additional examples primarily reinforce existing patterns rather than introducing new ones, raising the risk of overfitting to synthetic pronunciations. Third, the domain gap between synthetic and real speech may become more prominent as synthetic speech data dominates the training phase, reducing the effectiveness of real speech data fine-tuning in correcting artifacts.

Taken together, these findings suggest that an approximately 2:1 synthetic-to-real ratio (as implemented in Model III) achieved the optimal balance, maximizing tonal clarity and phonetic robustness while preserving natural human prosody.

VII. STATISTICAL VALIDATION

A. Inter-Rater Reliability

While a pool of 10 native speaker evaluators represents a compact sample size, recruiting qualified fluent speakers of an endangered, unwritten dialect presents acute resource constraints. To ensure the statistical reliability of this sample, the Kappa statistic [42] was utilized to measure the degree of agreement among the 10 native speaker evaluators (inter-rater agreement). Kappa values range from -1 to +1, where a score of +1 indicates perfect agreement and 0 indicates agreement purely by chance. The inter-rater agreement is calculated using the following formula:

$$K = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where, P_o (observed agreement) represents the proportion of instances where evaluators assigned the same rating to a speech sample, while P_e (agreement expected by chance) accounts for the likelihood that evaluators would agree on the rating for a speech sample purely by chance.

The 10 native speaker evaluators participated in a calibration phase, where they rated each other's Penang Hokkien speech on a 5-point scale (1 = poor, 5 = excellent). The resulting Kappa score of 0.71 indicated substantial agreement based on the established interpretation scale for inter-rater reliability [43]. This high degree of consensus

confirmed that the evaluators shared a consistent and reliable standard for linguistic authenticity in Penang Hokkien, ensuring that the Mean Opinion Score (MOS) ratings provided in the subjective evaluation are both stable and reliable.

B. Statistical Significance of MOS Differences

To determine whether differences in perceived speech quality among all evaluated systems were statistically meaningful, a one-way ANOVA [44] was conducted across the MOS scores for all model configurations. The ANOVA revealed a significant main effect ($p < 0.05$), indicating that at least one system differed significantly from the others.

Post-hoc pairwise comparisons (Tukey's HSD [45]) were then performed to identify which specific model pairs exhibited statistically significant differences. Among the significant comparisons, Model III (MOS = 3.92) demonstrated a statistically significant improvement over the Baseline Model I trained solely on the 45-minute real dataset (MOS = 3.57), with $p < 0.05$. Significant differences were also observed between Model III and several other lower-performing augmentation variants, confirming the robustness of the proposed method across multiple comparative conditions.

VIII. Limitation

Despite the successful outcomes, this research has several limitations. The primary limitation is the use of a single-speaker, 45-minute real-speech dataset. This constraint reflects practical realities, limited funding, limited speaker availability, and the absence of any larger curated Penang Hokkien corpora. As such, the dataset used here represents one of the most complete resources available for this critically low-resource dialect. The motivation for proceeding with such a small corpus was to conduct a feasibility study demonstrating what is realistically achievable for endangered or under-documented languages under true data scarcity. Nevertheless, the limited dataset constrains the model's ability to generalize across diverse prosodic patterns and emotional expressions, and this gap contributes to the remaining MOS deficit relative to groundtruth speech. Second, standard automated metrics such as Word Error Rate (WER) or Character Error Rate (CER) via a standardized Automatic Speech Recognition (ASR) model were not viable due to the non-existence of a baseline ASR system for this dialect. Consequently, spectral fidelity via MCD and perceptual tracking via multi-dimensional MOS were prioritized.

IX. Conclusion

This study successfully trained the first Text-to-Speech model for Penang Hokkien, addressing the critical challenges of phonological sparsity and resource scarcity inherent in unwritten tonal dialects. By implementing a two-stage fine-tuning approach using the SpeechT5_TTS model architecture, it was demonstrated that a phonetically-rich foundation, constructed from a syllable-level concatenative synthetic corpus covering comprehensive syllable-tone combinations, is essential for achieving high-fidelity synthesis and tonal accuracy. Technical refinements, including a 600-ms cross-fading technique to mitigate boundary artifacts and the use of numerical tone markers to reduce token sparsity, contributed to an overall Mean Opinion Score (MOS) of 3.92. Further cross-language validation is required to definitively generalize these findings to other tonal families. We position this work as an initial case study and a technological blueprint for the digital preservation of endangered, unwritten tonal dialects.

The synthesized speech samples can be found at https://cclia19.github.io/penanghokkientts/.

Acknowledgment

This study was supported by the UTAR Research Fund (IPSR/RMC/UTARRF/2022-C1/J01), Universiti Tunku Rahman, Malaysia.

References

[1] H. T. Ling, "Penang Hokkien Speech Analysis," Medium. 2021. [Online] Available: https://linghuiting.medium.com/penang-hokkien-speech-analysis-329f20248bb.

[2] P. Nambiar, "Hokkien on its last legs, warns linguist", Free Malaysia Today (FMT), 2020. [Online] Available: https://www.freemalaysiatoday.com/category/leisure/2020/08/04/hokkien-on-its-last-legs-warns-linguist/ (Accessed: 19 Jan 2024)

[3] J. Ao, et al., "SpeechT5: Unified-Modal Encoder-Decoder Pre-training for Spoken Language Processing", arXiv preprint arXiv:2110.07205, 2021.

[4] H. Zen, et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech", in Interspeech, Graz, Austria, 2019, pp. 1526-1530.

[5] F. S. Oliveira, et al., "CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages", in International Conference on Text, Speech, and Dialogue, Cham: Springer Nature Switzerland, 2023. pp. 188-199.

[6] A. W. Black, "Cmu wilderness multilingual speech dataset", in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5971-5975.

[7] J. Meyer, et al., "BibleTTS: a large, highfidelity, multilingual, and uniquely African speech corpus", in Interspeech, Incheon, Korea, 2022. pp. 2383-2387.

[8] A. Debnath, et al., "Low-Resource End-to-end Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning", in 2020 IEEE 17th India Council International Conference (INDICON), NSUT, Delhi, India, 2020, pp. 1-5.

[9] E. Cooper, "Text-to-speech synthesis using found data for low-resource languages", Ph.D dissertation, Columbia University, New York, US, 2019.

[10] A. Amalas, et al., "A multilingual training strategy for low resource Text to Speech", arXiv preprint arXiv:2409.01217, 2024.

[11] M. Baali, et al., "Unsupervised data selection for TTS: using Arabic Broadcast News as a case study", arXiv preprint arXiv:2301.09099, 2023.

[12] W. Fang, Y. A. Chung, and J. Glass, "Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models", arXiv preprint arXiv:1906.07307, 2019.

[13] Y. Jia, et al., "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS", in Interspeech, Brno, Czechia, 2021, pp. 151-155.

[14] A. H. Liu, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning", in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7259-7263.

[15] H. Zhang and Y. Lin, "Unsupervised Learning For Sequence-to-sequence Text-to-speech For Lowresource Languages", in Interspeech, Shanghai, China 2020, pp. 3161-3165.

[16] K. Lakhotia, et al., "On generative spoken language modeling from raw audio", Transactions of the Association for Computational Linguistics, vol. 9, 2021, pp. 1336-1354.

[17] J. Ni, et al., "Unsupervised Text-to-Speech Synthesis by Unsupervised Automatic Speech Recognition", in Interspeech, Incheon, Korea, 2022, pp. 461-465.

[18] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning", in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 2017, pp. 301-308.

[19] Y. Ren, et al., "Almost unsupervised text to speech and automatic speech recognition", in International Conference on Machine Learning, Long Beach, CA, USA, 2019, pp. 5410-5419.

[20] Z. Byambadorj, et al., "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, 2021, pp. 1-20.

[21] K. K. Lakshminarayana, et al., "Low-Resource Text-to-Speech Synthesis Using Noise-Augmented Training of ForwardTacotron," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1-5.

[22] T. Tao, et al., "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning", in Interspeech, Graz, Austria, 2019. pp. 2075-2079.

[23] J. Xu, et al., "Lrspeech: Extremely low-resource speech synthesis and recognition", in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020. pp. 2802-2812.

[24] Y. Zhang, et al., "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning", in Interspeech, Graz, Austria, 2019, pp. 2080-2084.

[25] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages", IEEE Access, 2020, Vol. 8, pp. 179798-1798812.

[26] C. H. Tan, Penang Hokkien for Penangites and Tourists. MPH Group Publishing Sdn Bhd, 2014.

[27] Learn Penang Hokkien. [Facebook]. Available at: https://www.facebook.com/groups/learnpenanghokkien/

[28] Y. L. Lai, J. Y. M. Khaw, S. P. Lim, and T. P. Tan, "Text Normalisation of Penang Hokkien Dialect Leveraging Adapted Soundex Algorithm," 2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS), Bangkok, Thailand, pp. 48-54, Sep. 2024

[29] Y. Tabet, A. Boustil, and K. Baiche, "Speech Analysis-Synthesis using Deep Neural Networks: A Review", in The 1st National Workshop on Wireless Network, Cloud Computing and Cryptography (WWN3C'2023), Boumerdes, Algeria, 2023, pp. 110-114.

[30] A. Franck, "Efficient Algorithms and Structures for Fractional Delay Filtering Based on Lagrange Interpolation", Journal of the Audio Engineering Society, vol. 56, no. 12, 2009, pp. 1036-1056.

[31] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis", Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 17022-17033.

[32] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment", in Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1, 1993, pp. 125–128.

[33] ITU-T Recommendation P.85, "A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices", International Telecommunication Union, 1994.

[34] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech : Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," Ph.D. dissertation, Université Paris Sud-Paris XI; Institut Polytechnique (Hanoï), 2015.

[35] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 9, 2015, pp. 1469-1477.

[36] SoX: Sound eXchange audio manipulation tool. 2024. [Online] Available: http://sox.sourceforge.net.

[37] T. Ko, et al., "Audio augmentation for speech recognition," in Interspeech, Dresden, Germany, 2015, pp. 3587-3589.

[38] A. Polyak, et al., "Speech resynthesis from discrete disentangled self-supervised representations," in Interspeech, Brno, Czechia, 2021, pp. 3615-3619.

[39] J. Tang and M. Li, "End-to-End Mandarin Tone Classification with Short Term Context Information," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 878-883.

[40] J. Vepa, S. King, and P. Taylor, "New Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 223-226.

[41] J. Zhong, et al., "Pairwise Evaluation of Accent Similarity in Speech Synthesis," *arXiv preprint arXiv:2505.14410*, 2025.

[42] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability", Educational and Psychological Measurement, vol. 33, no. 3, 1973, pp. 613–619.

[43] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol.33, no.1, 1977, pp. 159-174.

[44] Fisher, R. A., Statistical Methods for Research Workers. Oliver and Boyd, 1925.

[45] Tukey, J. W., "Comparing individual means in the analysis of variance," Biometrics, vol. 5, no. 2, pp. 99-114, Jun. 1949.

Exploring New Possibilities: Virtual Reality in Supporting Children with Autism

Brayan J. Lozano-Martel, Eduardo F. Araujo-Vasquez, Segundo E. Cieza-Mostacero
Cesar Vallejo University, Trujillo, Peru

***Abstract*—Children with autism spectrum disorder (ASD) often require structured opportunities to practice social interaction, following instructions, and recognizing emotions in safe, controlled environments. This study presents and evaluates a mobile virtual reality (VR) application designed to train these skills through simulated everyday scenarios. An applied quantitative study with an experimental design involving control and experimental groups was conducted with 10 children with ASD from the Sagrada Familia Special Basic Education Center (Trujillo, Peru). The experimental group (n=5) used a Unity 3D/C# VR application featuring four scenarios (park, store, pedestrian crossing, and emotion recognition), while the control group (n=5) followed conventional activities. Outcomes were collected through direct observation (60 records; 30 per group) and analyzed statistically, with findings interpreted cautiously due to the repeated-observation structure of the data. The post-test results showed statistically significant differences between groups, with the experimental group performing better on all evaluated indicators (social interaction, following instructions, emotion recognition, and average time to detect oncoming traffic). However, due to the small sample size and repeated observations, these findings should be interpreted as preliminary evidence. The results suggest that mobile VR may complement traditional interventions to strengthen social and emotional skills in children with ASD.**

Keywords*—*Autism Spectrum Disorder; virtual reality; mobile application; social skills training; immersive learning; special education

I. Introduction

In Peru, between 2020 and 2023, there has been a sustained increase in diagnoses of autism spectrum disorder, reflected in greater demand for specialized health and education services. National reports indicate that the highest concentration of diagnosed cases corresponds to school-age children, which has created a significant challenge for public institutions responsible for providing educational and therapeutic care, especially in contexts where access to specialized technological resources is still limited [1]. Although advances in diagnosis have enabled earlier detection, significant gaps in access to appropriate interventions and support services remain, highlighting the lack of comprehensive policies to accompany people with autism throughout their lives [2].

In the United States, epidemiological studies conducted by national monitoring networks indicate that autism spectrum disorder affects approximately 1.7% of children, equivalent to 1 in 59 school-age children. This prevalence has been identified from systematic assessments conducted in multiple states, demonstrating a significant impact of ASD on health and education systems, as well as the need for early interventions and specialized support strategies [3]. In Peru, around 1% of the population is on the spectrum, and more than 70% of registered cases correspond to children under the age of eleven, reinforcing the need to strengthen educational and technological support programs in special education [4].

Recent meta-analyses indicate that immersive virtual reality-based interventions can improve cognitive, social, and emotional skills in children and adolescents with ASD, providing structured, safe, and engaging practice opportunities [5]. Furthermore, virtual reality applications designed for high-functioning children with autism can effectively support social skill development, facilitate engagement in controlled, immersive environments, and help mitigate difficulties in social interaction and following structured instructions [6].

These difficulties can affect their participation in daily activities, limit their autonomy, and generate anxiety or discomfort in new situations. In educational and social contexts, many children with autism have problems relating to their peers, following instructions in public spaces, or identifying risky situations, such as when crossing a street. This highlights the need for safe, controlled, and personalized teaching methodologies that allow for the progressive development of social skills in structured environments [7].

In the city of Trujillo, the Sagrada Familia school, a Special Basic Education Center (CEBE), serves children diagnosed with ASD and focuses its work on the development of social and emotional skills. However, it faces challenges due to the limited availability of technological tools adapted to the teaching-learning process, which affects the behavioral progress and autonomy of students [7].

In this scenario, virtual reality (VR) emerges as a technology capable of offering immersive, safe, and personalized environments that facilitate learning through the simulation of everyday experiences. Recent studies show that the use of virtual environments promotes motivation, concentration, and emotional understanding, contributing significantly to the development of social skills in children with autism spectrum disorder [8].

In response to this problem, a mobile virtual reality application based on social environments was designed and developed with Unity 3D and the C# programming language. The application integrates everyday scenarios such as a park, a store, a crosswalk, and emotional recognition activities, aimed at promoting social interaction, decision-making, and understanding of emotions in an educational and therapeutic context [9].

The application was developed using the SUM methodology, which proposes a structured and modular approach to the construction of educational software, allowing iterative adjustments to be made based on observations and the performance of the participating children. This methodology facilitated the progressive adaptation of virtual environments to the needs of the intervention process [10], [11].

The main objective of this research was to improve social skills in children with autism spectrum disorder through the implementation of a virtual reality mobile application. Specifically, the aim was to increase the proportion of social interaction in groups of people, increase the frequency of following instructions in everyday environments such as shops, decrease the average time to detect oncoming traffic before crossing the street, and increase the emotion recognition rate in everyday situations.

Research hypotheses. The general hypothesis states that the implementation of a virtual reality mobile application based on social environments significantly improves social skills in children with autism spectrum disorder. The following specific hypotheses were also proposed: (H1) the use of the application increases the proportion of social interaction; (H2) it increases the frequency of following instructions in everyday environments; (H3) it decreases the average time to detect oncoming traffic before crossing the street; and (H4) it increases the emotion recognition rate in everyday situations.

II. Theoretical Framework

A. Virtual Reality

Virtual reality (VR) is an immersive technology that allows the user to enter a computer-generated three-dimensional environment, in which they can interact with objects, characters, or simulated situations in real time using specialized headsets and controllers. This technology is characterized by immersion, which generates a sense of presence in the virtual environment; interactivity, which allows the user to actively influence the development of the experience; and the ability to simulate controlled and safe contexts, facilitating the repetitive practice of skills without real risks. VR also offers immediate feedback and personalization of the environment, aspects that favor experiential learning and progressive adaptation by the user, especially in educational and therapeutic contexts [12].

B. Social Skills

Social skills are a set of learned behaviors that allow people to interact appropriately with their environment, express emotions, establish interpersonal bonds, and comply with social norms. These skills involve the ability to communicate, cooperate, empathize, and resolve conflicts in different contexts [5].

In the case of children with ASD, limitations in verbal and nonverbal communication hinder the acquisition of these skills, affecting their school and social integration. The development of social skills requires structured environments and methodologies that encourage observation, imitation, and constant practice of appropriate behaviors [6].

C. SUM Methodology

The SUM (Scrum Unified Methodology) methodology is considered an agile software development model that combines the iterative principles of Scrum with a structured approach to software development. This approach prioritizes continuous collaboration between the work team and the end user, encouraging partial deliveries, constant feedback, and adaptability to change [10]. SUM allows for the incremental planning and development of technological products, ensuring quality, flexibility, and control of time and costs. This methodology is suitable for educational projects, as it integrates stages of design, validation, and continuous improvement according to user needs [11].

Together, virtual reality as an immersive technology, the development of social skills as an intervention variable, and the SUM methodology as an agile development approach form the theoretical framework that underpins the design, implementation, and evaluation of the mobile application proposed in this research, aimed at improving social behavior in children with autism spectrum disorder.

III. Materials and Methods

A. Type of Research

This was an applied study, as it focused on the implementation of a social environment-based virtual reality mobile application with the aim of improving the behavior of children diagnosed with autism spectrum disorder (ASD). This type of study seeks to generate practical and tangible solutions using scientific knowledge [13]. Likewise, a quantitative approach was used, since the data obtained were numerical and were analyzed using statistical techniques to verify the validity of the hypotheses proposed [14].

B. Research Design

A true experimental design was used, aimed at determining the effectiveness of the virtual reality mobile application in improving the behavior and social skills of children with ASD. This design is characterized by the deliberate manipulation of an independent variable and the observation of its effects on a dependent variable under controlled conditions. Experimental studies allow causal relationships between variables to be established through the use of control and experimental groups, ensuring the internal validity and reliability of the results [14].

In this parallel-group true experimental design, the Control Group served as the concurrent behavioral reference, reflecting performance under conventional educational conditions without VR exposure. Both groups were matched at allocation by age, educational level, and ASD diagnosis profile, supporting initial group equivalence. Accordingly, between-group post-test comparisons provide the primary evidence of the intervention's effect, consistent with parallel-group experimental designs in which group equivalence is established at allocation [15].

Because detailed pretest values were not reported in the manuscript, full statistical verification of baseline equivalence is limited.

To reduce this limitation, participants were allocated considering similarity in age, educational level, and confirmed ASD diagnosis. The Control Group served as a concurrent behavioral reference under conventional educational conditions. Therefore, the results should be interpreted as preliminary evidence [14].

C. Variables and Operationalization

1) Independent variable: Virtual reality.

2) Conceptual definition: It is a technology that allows the creation of three-dimensional environments with which users can interact in an immersive way [16].

3) Operational definition: The virtual reality application was used by the Experimental Group. A nominal scale was used.

4) Dependent variable: Behavior of children with autism.

5) Conceptual definition: The behavior of children with autism is defined as qualitative alterations in social interaction, communication, and repetitive and restricted behavior patterns [5].

6) Operational definition: The variable Behavior of children with autism was measured using four indicators: proportion of social interaction, frequency of following instructions, average time to detect oncoming traffic before crossing, and emotion recognition rate. All indicators were measured on a ratio scale.

D. Population, Sample, and Sampling

The population consisted of children diagnosed with autism spectrum disorder (ASD) attending the Sagrada Familia Special Basic Education Center in Trujillo, Peru. An intentional sample of ten (10) participants was selected and divided into two groups with similar characteristics: five in the experimental group and five in the control group, maintaining similarity in age, educational level, and behavioral characteristics. The selection was made considering inclusion criteria such as confirmed diagnosis of ASD, regular attendance at the center, and informed consent from parents or guardians [17].

Due to limitations in parental authorization and the small number of participants, the study included a total of 10 children with ASD (5 per group). While inferential statistical tests, including the Mann–Whitney U test, were applied to compare outcomes between groups, it is important to recognize that small sample sizes can produce unstable effect estimates. No formal statistical power analysis was conducted due to the limited sample size; therefore, the results should be interpreted as preliminary evidence rather than definitive conclusions. To strengthen behavioral documentation, 30 observational records per group were collected through repeated observations during the intervention sessions. These repeated measures provide insight into the children's performance but do not increase the statistical power at the participant level.

During data collection, a total of 60 observational records were obtained (30 per group), corresponding to repeated observations of the 10 participants. Although participants constituted the primary sampling unit, the observational records were used to document behavioral performance across intervention sessions. Therefore, the inferential results should be interpreted as exploratory and preliminary evidence rather than definitive participant-level effects [18].

E. Data Collection Techniques and Instruments

Data collection was carried out through direct observation, a technique that allowed for real-time recording of the children's behaviors and responses during the intervention sessions. This method provided accurate and objective information about the participants' actions in their controlled educational environment, ensuring the reliability of the results [18].

The instrument used was an observation form validated by expert review, designed to systematically document the indicators defined in the dependent variable: social interaction, following instructions, traffic identification, and understanding emotions.

The observation instrument underwent a formal content validation process conducted by a qualified expert in autism spectrum disorder and special education. The validator held expertise in behavioral assessment and therapeutic intervention with ASD populations, and evaluated the instrument across multiple dimensions: 1) clarity, assessing whether each item was written in clear, unambiguous language appropriate for the observational context; 2) relevance, confirming that each indicator was directly aligned with the behavioral dimensions targeted by the intervention (social interaction, instruction-following, traffic hazard detection, and emotion recognition); 3) coherence, verifying that the indicators were internally consistent with the theoretical framework underpinning the study; 4) pertinence, ensuring that each item was appropriate and necessary for measuring the behavioral outcomes in children with ASD in educational settings; and 5) sufficiency, confirming that the set of indicators as a whole provided adequate coverage of the dependent variable. Based on the expert's review, minor wording adjustments were made to two items to improve observational objectivity, and the instrument was confirmed as suitable for measuring the target behavioral indicators in this population. The expert's assessment provided evidence of content validity, supporting the use of the observation form for the purposes of this study. However, because the instrument was reviewed by a single expert and no inter-rater reliability or agreement coefficient was calculated, this aspect is acknowledged as a methodological limitation.

The data obtained were processed using Jamovi software version 2.3.28, applying descriptive and inferential statistics. For group comparisons, the Shapiro–Wilk test was used to assess normality; Student's t-test was applied to normally distributed data, and the Mann–Whitney U test was used otherwise [15].

F. Procedures

The study was carried out in three stages following the SUM methodology, which is suitable for the agile and incremental development of educational software. In the first stage, the virtual reality mobile application was designed and programmed using the Unity 3D platform and the C# language, integrating three-dimensional environments focused on improving the social behavior of children with autism spectrum

disorder (ASD). Subsequently, functionality, usability, and performance tests were carried out to ensure the stability of the system before its application in the educational environment.

During the second stage, participants in the Experimental Group used the mobile application through Oculus Quest 2 virtual reality headsets in controlled sessions. The activities included interactive scenarios representing everyday situations, such as interacting in a park, shopping, crossing a street, and recognizing emotional expressions. Each session lasted approximately 15 to 20 minutes and was conducted under the supervision of educational staff. To reduce potential sensory burden, participants were allowed to pause or stop the activity if signs of discomfort, fatigue, anxiety, visual overstimulation, or auditory sensitivity appeared. In addition, the virtual scenarios were designed with simple visual elements, guided interactions, and limited exposure time to minimize sensory overload during the intervention. At the same time, the Control Group continued with conventional educational activities without VR exposure.

In the third stage, the observation form was administered to both groups to document behavioral performance across the four indicators: social interaction, following instructions, average time to identify traffic, and emotion recognition. For the Experimental Group, this corresponded to the post-intervention phase; for the Control Group, it captured concurrent performance under conventional educational activities, serving as the parallel behavioral reference. Data were then processed and analyzed to determine the effect of the VR application.

IV. Software Methodology

The virtual reality mobile application was developed using the Scrum Unified Methodology (SUM), an agile approach geared toward the development of interactive systems and immersive applications. This methodology was selected for its iterative and flexible nature and its suitability for the development of virtual reality-based educational applications, allowing for continuous improvements based on feedback and system testing.

The SUM methodology is structured in five phases: Conception, Planning, Initial Development, Beta Version, and Closure, which allowed for the progressive organization of the design, implementation, and validation of the application aimed at improving the social behavior of children with autism spectrum disorder (ASD).

A. Platform and Tools

The virtual reality application was implemented for the Oculus Quest 2 platform, with the aim of providing an autonomous immersive experience for sessions in the educational environment. Development was carried out with Unity 2022.3.16f1 and programming in C#; the environments and characters were integrated using 3D models from asset catalogs. User interaction (selection, button activation, and object manipulation) was implemented with XR Interaction Toolkit under the OpenXR standard. For the recording and consolidation of results per session, a cloud service was integrated using Supabase, consumed from Unity with UnityWebRequest.

To run the application on the headset, the project was configured in Unity for Android, generating an APK package compatible with (Oculus/Meta) Quest 2. The compilation was performed using the Android SDK and OpenJDK, and the APK was then installed and managed on the device using Meta Quest Developer Hub, allowing controlled tests to be run in a real environment. During these tests, the correct loading of scenes, interaction via controllers, and network connectivity for sending records to the database (Supabase) were verified, ensuring the storage of results per session.

B. Software Architecture

The solution was organized in a modular fashion: 1) a navigation module for accessing the objectives/scenarios, 2) an XR interaction module for user selection and control, 3) logic modules per scenario (park, store, intersection, and emotional recognition) with decision and timing controllers, and 4) a logging module that stores user responses/times and manages the sending of results. This structure facilitated iterative adjustments during development and stabilization before the experimental phase. As shown in Fig. 1, the XR application was developed in Unity using C# for deployment on Meta Quest 2, relying on the Android toolchain (Android SDK/OpenJDK) and Supabase as the database backend to store user logs and support result transmission.

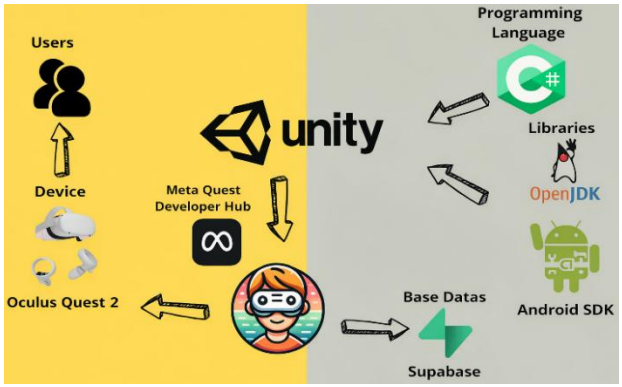


Fig. 1. Software architecture.

C. Scenario Implementation and Recording

The application includes four objectives: 1) social interaction in the park, where the child practices greetings and interaction with groups of people and the system counts successful interactions; 2) interactive store, where the child selects products from a list and the system counts correct selections; 3) pedestrian crossing/risk situation, where the child crosses while avoiding vehicles and attempts and time are recorded; and 4) emotional recognition, where the child identifies emotions on faces and the success rate is recorded. In all cases, the system saves the results at the end of each activity for later analysis.

D. Testing and Debugging

Before the intervention, interaction tests and iterative corrections were performed to ensure system stability. Among the incidents resolved, the interaction of panels/buttons in VR was corrected by adjusting the raycaster, and the sending of results to the database was stabilized by configuration and connection changes in OpenXR.

E. Conception Phase

In the conception phase, an analysis was carried out of the educational context and behavioral needs of children diagnosed with autism spectrum disorder. This stage helped identify the main difficulties to be addressed, such as social interaction, following instructions, recognizing emotions, and identifying risky situations in everyday environments.

Based on this analysis, the overall objective of the application was defined, which was to provide a safe and controlled virtual environment that allows children to practice social skills through immersive experiences. Likewise, the functional scope of the system was established, and the profile of the end users was determined, prioritizing accessibility and ease of use.

Based on the analysis carried out in this phase, the needs and objectives of the application were identified, which are summarized in Table I.

TABLE I. NEEDS ANALYSIS IN THE CONCEPTION PHASE

Aspect Analyzed	*Description*
Context of the problem	Children diagnosed with autism spectrum disorder have difficulties with social interaction, following instructions, recognizing emotions, and understanding risky situations in everyday environments.
Target user	Children with autism spectrum disorder (ASD) attending a special elementary school.
Identified need	To have a technological tool that allows social skills to be practiced in a safe, controlled, and repeatable environment.
Objective of the application	To provide immersive virtual environments that promote the development of social and behavioral skills.
Solution strategy	Implementation of a virtual reality mobile application based on interactive social scenarios.
Proposed environments	Park, store, crosswalk, and emotional recognition activities.

F. Planning Phase

During the planning phase, the functional and non-functional requirements of the application were defined. The general structure of the system was designed, establishing the different interactive virtual environments, which included scenarios such as a park, a store, a pedestrian crossing, and emotional recognition activities.

Likewise, the technological tools for software development were selected, using Unity 3D as the development engine and C# as the programming language. The project activities were organized according to development priorities, and a schedule was established that allowed for the progressive implementation of the functionalities.

As a result of the analysis carried out in the conception phase, the functional requirements of the virtual reality application were defined in the planning phase. These requirements made it possible to establish the functionalities necessary for the development of social scenarios and to guarantee an adequate user experience for children with autism spectrum disorder. Table II presents the functional requirements considered for the design of the system.

TABLE II. FUNCTIONAL REQUIREMENTS OF THE VIRTUAL REALITY APPLICATION

ID	*Requirement*
RF1	The application must provide immersive virtual reality environments geared toward the development of social skills.
RF2	The system must allow the user to interact with virtual objects and characters within each scenario.
RF3	The application must include everyday social scenarios such as a park, a store, and a crosswalk.
RF4	The system must incorporate activities for recognizing and understanding emotions.
RF5	The application must allow for simple and intuitive navigation between the different scenarios.
RF6	The system must record user interactions and responses during usage sessions.
RF7	The application must be suitable for children with autism spectrum disorder, prioritizing visual simplicity and ease of use.

G. Initial Development Phase

In this phase, the main functionalities of the virtual reality application were implemented. Three-dimensional environments were developed and integrated in Unity 3D, programming the interactive elements that allow user navigation, interaction with virtual objects and characters, as well as the execution of guided activities.

In addition, the system logic was implemented to record user responses and behaviors during usage sessions. Iterative testing was performed to verify system stability, usability, and proper interaction between components, allowing for early adjustments to the application's design and operation.

As part of the initial development, the application was structured into different virtual social scenarios, each aimed at reinforcing a specific skill. Table III presents the relationship between the scenarios developed and the social skills addressed in each one.

TABLE III. VIRTUAL SCENARIOS DEVELOPED AND SKILLS ADDRESSED

Virtual Scenario	Social Skill Addressed
Park	Social interaction in public spaces
Store	Following instructions in everyday environments
Crosswalk	Identification of risky situations
Recognition of emotions	Understanding emotions in social situations

H. Beta Version Phase

The beta version phase was aimed at validating the application in a controlled educational environment. The system was implemented using virtual reality headsets, allowing children in experimental groups to interact with the different scenarios during scheduled sessions.

During this stage, feedback was collected from direct observation and support from educational staff, which allowed for the identification of opportunities for improvement related to visual clarity, interaction time, and navigation within the virtual environment. Based on these results, minor adjustments were made to optimize the user experience and reduce possible sensory overload.

To validate the application's performance in a controlled environment, a beta version of the virtual reality system was developed. This version allowed for the evaluation of the correct execution of the virtual scenarios, user interaction with the elements of the environment, and the overall stability of the application before its experimental implementation. As shown in Fig. 2, the beta version enabled verification of scenario execution, user interaction with virtual elements, and overall application stability.



Fig. 2. Virtual scenario of the virtual reality application in execution.

I. Closing Phase

In the closing phase, the final version of the virtual reality mobile application was consolidated. The correct functioning of all modules was verified, and compliance with the objectives established in the initial phases was checked.

Finally, the application was prepared for implementation in the experimental phase of the research, serving as the main tool for data collection and evaluation of the impact of virtual reality on improving the social behavior of children with autism spectrum disorder.

Upon completion of the application's development, the final components of the system were consolidated, ensuring its proper functioning and alignment with the objectives set forth. Table IV summarizes the main components that make up the final version of the virtual reality application.

TABLE IV. FINAL COMPONENTS OF THE VIRTUAL REALITY APPLICATION

Component	Description
Development engine	Unity 3D
Programming language	C#
Virtual scenarios	Park, store, crosswalk, and recognition of emotions
Interaction	Navigation and interaction with virtual objects and characters
Platform	Mobile application with virtual reality viewer
Final status	Functional system ready for experimental implementation

V. RESULTS

A. Descriptive Analysis

The descriptive analysis of the observational results showed differences between the Experimental Group (EG) and the Control Group (CG) across the four evaluated indicators. The CG results reflect behavioral performance under conventional educational activities across the observation period, serving as the parallel baseline reference for this experimental design. Higher values indicate better performance for social interaction, following instructions, and emotion recognition, whereas a lower value indicates better performance for average traffic detection time. The descriptive results are presented in Table V.

TABLE V. DESCRIPTIVE RESULTS OF SOCIAL AND EMOTIONAL INDICATORS IN CONTROL AND EXPERIMENTAL GROUPS

Indicator	Control Group (CG)	Experimental Group (EG)	Interpretation
Proportion of social interaction (%)	35.47 ± 17.25	54.29 ± 23.89	Higher is better
Frequency of following instructions (%)	26.61 ± 21.27	63.29 ± 26.05	Higher is better
Emotion recognition rate (%)	26.67 ± 14.58	73.33 ± 19.62	Higher is better
Average time to detect oncoming traffic (s)	45.71 ± 2.02	24.29 ± 2.46	Lower is better

The Experimental Group obtained higher average values than the Control Group in the proportion of social interaction, frequency of following instructions, and emotion recognition rate. Specifically, the Experimental Group reached 54.29% in social interaction, 63.29% in following instructions, and 73.33% in emotion recognition, whereas the Control Group obtained 35.47%, 26.61%, and 26.67%, respectively. For the average time to detect oncoming traffic before crossing, the Experimental Group recorded a lower mean time of 24.29 seconds compared with 45.71 seconds in the Control Group. Since lower response time indicates better performance in this indicator, this result suggests that the Experimental Group may detect risky traffic situations faster, providing preliminary evidence of the effect of the VR application.

B. Inferential Analysis

Table VI presents the results of the inferential analysis performed for the four indicators evaluated in the research, with the aim of determining whether there are statistically significant differences between the Control Group (CG) and the Experimental Group (EG) after the implementation of the virtual reality mobile application.

According to the normality test results, nonparametric Mann–Whitney U tests were applied to the indicators of proportion of social interaction, frequency of following instructions, and emotion recognition rate, while Student's t-test was used for the indicator of average time to detect oncoming traffic before crossing.

The significance values obtained ($p < 0.05$) show statistically significant differences between the two groups in the four indicators analyzed. In particular, the results suggest that the Experimental Group performed better on indicators related to social and emotional skills, as well as a reduction in

average time to identify traffic, providing preliminary evidence of the effect of the VR application.

TABLE VI. INFERENTIAL STATISTICAL TESTS BY INDICATOR

Indicator	Test applied	Statistic	P-value	Interpretation
Proportion of social interaction	Mann–Whitney U	U = 267.50	p = 0.002	Significant difference
Frequency of following instructions	Mann–Whitney U	U = 145.00	p < 0.001	Significant difference
Average time to detect oncoming traffic before crossing	Student's t-test	t(58) = -36.87	p < 0.001	Significant difference
Emotion recognition rate	Mann–Whitney U	U = 37.00	p < 0.001	Significant difference

Inferential tests were calculated using observational records collected during the intervention sessions. Since these records correspond to repeated observations of 10 participants, the results should be interpreted as preliminary evidence rather than definitive participant-level effects.

VI. DISCUSSION

The study employed a parallel-group experimental design in which the Control Group (CG) served as the concurrent behavioral reference, documenting performance levels under conventional educational conditions without VR exposure across the same observation period (October 2–22, 2024). Both groups were matched at allocation by age, educational level, and confirmed ASD diagnosis, supporting comparability between groups before the experimental phase. The post-test results for the CG, therefore, represent the expected behavioral trajectory in the absence of the intervention, against which Experimental Group outcomes are compared. This design follows established practice in parallel-group experimental research, where between-group comparisons at post-test provide the primary evidence of intervention effect [15]. Findings should nonetheless be interpreted as preliminary evidence given the small sample size (n = 5 per group) and the use of repeated observational records.

For the first indicator (proportion of social interaction), the results showed a notable improvement in the experimental group compared to the control group after using the virtual reality mobile application. This increase is attributed to the fact that virtual environments allow social situations to be simulated in controlled contexts, facilitating the repetitive practice of social behaviors without the pressure of the real environment. Immersion and progressive interaction with virtual characters encouraged greater social participation among children with autism spectrum disorder, which is consistent with previous studies reporting improvements in social skills through interactive virtual environments [12]. In this context, social interaction is understood as the ability to initiate, maintain, and respond appropriately to relevant social exchanges [19].

For the second indicator (frequency of following instructions), the results showed a significant increase in the experimental group, which showed an improvement in the understanding and execution of instructions. This behavior can be explained by the structure of the virtual scenarios, which presented clear, sequential, and visually accessible orders, reducing distractions and facilitating sustained attention. Several studies indicate that structured virtual environments contribute to reducing anxiety and disruptive behavior, promoting compliance with instructions in children with autism spectrum disorder [20]. Likewise, difficulties in following instructions are often related to attention and sensory processing challenges, which can be supported through structured virtual reality interventions designed for high-functioning children with autism [6].

For the third indicator (average time to detect oncoming traffic before crossing), the results showed a significant reduction in time in the experimental group. This finding indicates that repeated exposure to simulated road scenarios allowed participants to recognize relevant stimuli more quickly and efficiently, strengthening their ability to respond to risky situations. Virtual reality offers a safe environment for training skills related to decision-making and hazard identification without exposing users to real risks [21]. Likewise, controlled repetition of complex scenarios has been shown to promote learning and anticipation of critical situations [22].

For the fourth indicator (emotion recognition rate), the results showed considerable improvement in the experimental group after the intervention. This result suggests that virtual scenarios aimed at emotional recognition facilitated the identification and interpretation of facial expressions and social contexts. The visual representation of emotions in structured virtual environments reduces cognitive overload and promotes social-emotional learning in children with autism spectrum disorder [23]. In this sense, emotional understanding is defined as the ability to recognize, interpret, and respond appropriately to one's own and others' emotions, constituting a fundamental skill for effective social interaction [19].

Overall, the results suggest that the mobile virtual reality application was associated with improvements in social interaction, following instructions, emotion recognition, and traffic detection skills in children with autism spectrum disorder. These results provide preliminary evidence of the potential of VR interventions. However, the study is limited by the small sample size and repeated observations, which may affect the generalizability of the findings. Additionally, the use of the Oculus Quest 2 headset may impose a potential sensory burden, such as visual or auditory overstimulation, which should be considered in future research.

VII. CONCLUSION

This study provides preliminary evidence that a mobile virtual reality application may contribute to improving social interaction, following instructions, emotion recognition, and traffic detection skills in children with autism spectrum disorder (ASD) within a controlled educational setting. Improvements were observed across all evaluated indicators, suggesting that VR-based interventions may serve as a valuable complement to traditional teaching methods for children with ASD.

These findings should be interpreted with caution due to the small sample size, the use of repeated observations, and the

potential sensory overload associated with the Oculus Quest 2 headset, which could influence engagement and performance during the intervention sessions.

Future research should include larger participant pools, explore additional educational contexts, and investigate the potential sensory effects of VR devices such as the Oculus Quest 2 to better understand their influence on engagement and performance.

REFERENCES

[1] G. Celis and M. Ochoa, «Trastorno del espectro autista (TEA) [Autism spectrum disorder]», in Revista de la Facultad de Medicina de la UNAM, vol. 65, n.º 1, pp. 77-171, enero-febrero 2022. doi: 10.22201/fm.24484865e.2022.65.1.02.

[2] V. Ruggieri, "El autismo a lo largo de la vida [Autism throughout life]," MEDICINA, vol. 82, Suppl. III, pp. 2–6, 2022, doi: 10.37610/MEDICINA.v82.supl.III.2022.

[3] A. Narzisi, Y. Alonso-Esteban, and F. Alcantud-Marín, «Autism and Children: Diagnosis, Functional Profiles and Intervention», Children, vol. 10, 2023, doi: 10.3390/children10030522.

[4] M. Baquerizo-Sedano, J. Lucero, and A. Taype-Rondan, "Autismo en Perú: estado actual [Autism in Peru: current status]," Rev. Cuerpo Med. HNAAA, vol. 16, no. 3, 2023, doi: 10.35434/rcmhnaaa.2023.163.2034.

[5] P. Mittal, M. Bhadania, N. Tondak, P. Ajmera, S. Yadav, A. Kukreti, S. Kalra, and P. Ajmera, «Effect of immersive virtual reality-based training on cognitive, social, and emotional skills in children and adolescents with autism spectrum disorder: A meta-analysis of randomized controlled trials», Research in Developmental Disabilities, vol. 151, pp. 104771, 2024. doi: 10.1016/j.ridd.2024.104771.

[6] M. Chiappini, C. Dei, E. Micheletti, E. Biffi, and F. A. Storm, «High-Functioning Autism and Virtual Reality Applications: A Scoping Review», Applied Sciences, vol. 14, n.º 7, pp. 3132, 2024. doi: 10.3390/app14073132.

[7] L. Liu, «Virtual Reality for Social Skills Training of Children and Adolescents with ASD: A Systematic Review», J. Educ. Hum. Soc. Sci., vol. 8, pp. 2061–2067, 2023, doi: 10.54097/ehss.v8i.4645.

[8] A. Frolli, G. Savarese, F. Di Carmine, A. Bosco, E. Saviano, A. Rega, M. Carotenuto, and M. C. Ricci, «Children on the Autism Spectrum and the Use of Virtual Reality for Supporting Social Skills», Children, vol. 9, 2022, doi: 10.3390/children9020181.

[9] C. Montalvo, Y. Novoa, and E. Nauca, "Aplicación móvil de realidad virtual para el aprendizaje estudiantil [Mobile virtual reality application for student learning]," Universidad de Lambayeque, 2020.

[10] N. Acerenza et al., "Una metodología para desarrollo de videojuegos: versión extendida [A methodology for video game development: extended version]," Technical Report RT 09-13, PEDECIBA, Instituto de Computación, Universidad de la República, Montevideo, Uruguay, 2009.

[11] J. G. Chero Caamaño, "Estudio de usabilidad de videojuegos web utilizando la metodología de desarrollo SUM [Usability study of web video games using the SUM development methodology]," B.S. thesis, Universidad Técnica de Machala, Machala, Ecuador, 2019.

[12] A. Soltiyeva, W. Oliveira, and M. Alimanova, «My Lovely Granny's Farm: An Immersive Virtual Reality Training System for Children with Autism Spectrum Disorder», Educ. Inf. Technol., vol. 28, pp. 16887-16907, 2023, doi: 10.1007/s10639-023-11862-x.

[13] G. Grebe, "Investigación científica e investigación aplicada en el Instituto de Salud Pública de Chile [Scientific and applied research at the Public Health Institute of Chile]," Rev. Inst. Salud Pública Chile, vol. 2, no. 1, pp. 37–43, 2018.

[14] R. Hernández-Sampieri and C. Mendoza, Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta [Research Methodology: Quantitative, Qualitative, and Mixed Approaches]. Mexico City, Mexico: McGraw-Hill Education, 2018, ISBN: 978-1-4562-6096-5.

[15] J. Zurita-Cruz, H. Márquez-González, G. Miranda-Novales, and M. Villasís-Keever, "Estudios experimentales: diseños de investigación para la evaluación de intervenciones en la clínica [Experimental studies: research designs for the evaluation of clinical interventions]," Rev. Alerg. Mex., vol. 65, no. 2, pp. 178–186, 2018, doi: 10.29262/ram.v65i2.376.

[16] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers and Education*, vol. 147, Article 103778, 2020, doi: 10.1016/j.compedu.2019.103778.

[17] A. Casteel, N. L. Bridier, «Describing Populations and Samples in Doctoral Student Research», International Journal of Doctoral Studies, vol. 16, pp. 339-362, 2021, doi: 10.28945/4766.

[18] C. Howell, W. Su, A. Nassel, A. Agne, and A. Cherrington, "Area based stratified random sampling using geospatial technology in a community-based survey," BMC Public Health, vol. 22, 2022, doi: 10.1186/s12889-022-13819-0.

[19] C. Berenguer, B. Roselló, and I. Baixauli, "Conductas de aprendizaje en niños con trastorno del espectro autista con y sin comorbilidad con déficit de atención e hiperactividad [Learning behaviors in children with autism spectrum disorder with and without comorbid attention deficit hyperactivity disorder]," *Int. J. Dev. Educ. Psychol.*, vol. 3, Spec. Issue 1, pp. 259–268, 2018, doi: 10.17060/ijodaep.2018.n1.v3.1269.

[20] R. Bossenbroek, A. Wols, J. Weerdmeester, A. Lichtwarck-Aschoff, I. Granic, and M. van Rooij, «Efficacy of a Virtual Reality Biofeedback Game (DEEP) to Reduce Anxiety and Disruptive Classroom Behavior: Single-Case Study», JMIR Ment. Health, vol. 7, n.o 3, 2020, doi: 10.2196/16066.

[21] I. Miller, B. Wiederhold, C. Miller, and M. Wiederhold, «Virtual Reality Air Travel Training with Children on the Autism Spectrum: A Preliminary Report», Cyberpsychol. Behav. Soc. Netw., vol. 23, n.o 1, 2019, doi: 10.1089/cyber.2019.0093.

[22] D. Johnston, H. Egermann, and G. Kearney, «SoundFields: A Virtual Reality Game Designed to Address Auditory Hypersensitivity in Individuals with Autism Spectrum Disorder», Appl. Sci., vol. 10, n.o 2996, 2020, doi: 10.3390/app10092996.

[23] J. Wu, Y. Sun, G. Zhang, Z. Zhou, and Z. Ren, «Virtual Reality-Assisted Cognitive Behavioral Therapy for Anxiety Disorders: A Systematic Review and Meta-Analysis», Front. Psychiatry, vol. 12, 2021, doi: 10.3389/fpsyt.2021.575094.

Augmented Reality in Automotive Technical Training: Technological Innovation Through Mobile Applications in Trujillo

An Immersive Technologies-Based Approach to Automotive Maintenance Training

Julissa B. Polo-Holguin, Eduardo F. Araujo-Vasquez, Segundo E. Cieza-Mostacero
Universidad César Vallejo, Trujillo, Perú

***Abstract*—This study evaluated the impact of an Augmented Reality (AR)-based mobile application on technical training in automotive maintenance at a company in Trujillo, Peru. The application was developed using Unity 3D, the Vuforia marker-based recognition engine, and C#, enabling technicians to interact with three-dimensional models of automotive components — including engines, valve trains, crankshafts, and braking systems — through touch-based rotation, zoom, and perspective controls overlaid on physical QR markers. This marker-based interaction architecture distinguishes the system by enabling step-by-step guided visualization of complex maintenance procedures without requiring physical equipment access. A posttest-only control group experimental design was employed with 60 technicians randomly assigned to experimental and control groups, using the Mobile-D methodology for software development. The results revealed a significant improvement in technical content comprehension, a 46.6% reduction in average training time (from 10,862 to 5,784 seconds), and increased satisfaction levels, reaching a mean score of 4.0 versus 2.6 in the control group. The Mann–Whitney U test confirmed statistically significant differences ($p = 0.001$) across all indicators, with large effect sizes (Rosenthal's $r \geq 0.86$). These findings suggest that AR-based mobile applications represent a viable instructional approach for automotive maintenance training, though broader generalization requires validation across diverse organizational contexts.**

***Keywords*—Augmented reality; marker-based AR; automotive maintenance training; mobile application; technical comprehension; training efficiency; experimental design**

I. Introduction

Digital transformation has profoundly reshaped training processes in technical education, particularly in fields that require the mastery of specialized practical skills, such as the automotive sector. Augmented Reality (AR) has emerged as an innovative technology with the potential to support the development of technical competencies by enabling the overlay of virtual content onto real-world environments through mobile devices. This approach enables technicians to explore simulations of complex systems and technical procedures without the need to manipulate actual physical equipment [1].

In the field of technical automotive education, the use of emerging technologies has become increasingly relevant due to the industry's dynamic nature and high technological demands. Advances in vehicle systems, electronic control, predictive maintenance, and automated process management have led educational institutions to rethink their teaching methodologies [2]. However, previous evidence in this area should be interpreted with caution. Although VR/AR-based educational environments have been proposed to support automotive and manufacturing training, some implementations have been developed mainly in university laboratory settings and evaluated through student experience or satisfaction measures rather than through controlled experimental comparisons in real workplace training contexts [3]. Therefore, while AR offers interactive virtual environments with potential to support safer and more accessible learning, further empirical evidence is still needed to evaluate its impact on measurable training outcomes such as technical comprehension, task completion time, and participant satisfaction in authentic automotive maintenance environments.

The contributions of Augmented Reality (AR) are not limited to the three-dimensional visualization of mechanical components, but also include step-by-step guided training through visual cues, animations, highlighted objects, and real-time instructional feedback. These features may support technician autonomy and reduce procedural errors, particularly in maintenance tasks that require spatial orientation and sequential execution. However, previous evidence suggests that AR effectiveness is task-dependent. For instance, one study found that traditional training performed better for single-level maintenance tasks, whereas AR showed greater advantages in complex multi-level tasks, although the generalizability of its findings was limited by specific hardware conditions, such as the restricted field of view of HoloLens 2, as well as by the need for further comparative and longitudinal evaluations [4]. Therefore, AR shows potential for practical maintenance training, but its effectiveness may vary according to task complexity and implementation conditions.

Several studies have reported that Augmented Reality (AR) can support industrial maintenance activities by providing contextual information, visual guidance, and interactive assistance during diagnostics, inspections, assembly, and maintenance procedures [5]. However, despite the promising results of AR applications in maintenance, their industrial adoption is still limited by technical constraints, fragmentation among hardware and software solutions, tracking accuracy issues, and the lack of maturity of some AR systems for real operational environments. Similarly, research on predictive

maintenance and intelligent sensors in smart factories highlights that efficiency improvements depend not only on visualization technologies, but also on the availability of reliable data, sensor interoperability, cybersecurity, and the effective integration of Industry 4.0 technologies [6].

At a global level, recent studies indicate that Augmented Reality (AR) is increasingly aligned with the principles of Industry 4.0, particularly in smart engineering manufacturing environments where technologies such as the Internet of Things (IoT), intelligent sensors, data analytics, and predictive maintenance are progressively integrated [7]. However, this alignment is often discussed from a technological or manufacturing perspective rather than from a pedagogical standpoint. Therefore, in automotive maintenance training, the value of AR should not be assumed only from its compatibility with Industry 4.0, but also from how its content, interaction design, and simulated procedures support the actual training needs of technicians.

In local contexts such as the city of Trujillo, where technical education programs seek to adapt to the challenges of a changing industrial environment, the incorporation of Augmented Reality (AR) may represent a strategic opportunity to strengthen the training of automotive technicians. The integration of mobile applications with AR content can provide more accessible training resources and reduce exclusive dependence on physical equipment; however, accessibility does not necessarily ensure effective learning, since outcomes also depend on instructional design, device availability, user familiarity with mobile technologies, and the realism of simulated procedures. Empirical evidence evaluating these benefits in local technical training settings remains scarce, underscoring the need for context-specific studies. Therefore, evaluating mobile AR in an automotive maintenance environment in Trujillo is necessary to determine whether these potential benefits translate into improvements in technicians' comprehension, training time, and satisfaction [8].

Technical training in the automotive field represents a fundamental pillar to ensure quality, efficiency, and safety in vehicle maintenance services. However, in contexts where traditional teaching methods still predominate—based on static theoretical lectures and limited instructional materials—training processes face significant barriers, resulting in a gap between the knowledge acquired in the classroom and the practical competencies required in the workplace, which affects the professional performance of future technicians. Additionally, the lack of innovative pedagogical strategies and the limited use of educational technologies hinder the development of technical, analytical, and problem-solving skills, which are essential to address the continuous technological advances in the automotive sector [9].

Difficulties in understanding complex mechanical procedures, excessive session duration, and low participant motivation reflect a mismatch between industry demands and the competencies acquired in the classroom. This situation becomes even more critical in the context of the rapid technological advancement of the automotive industry, where new vehicle models integrate electronic and digital systems that require more specialized technical skills [10].

One of the alternatives with the greatest potential has been the development of mobile applications based on Augmented Reality (AR), a technology that enables the visualization and interaction with three-dimensional models of automotive components through mobile devices [11]. This approach allows the simulation of real procedures without requiring direct access to physical equipment, which may reduce operational risks and facilitate autonomous training. However, its contribution to knowledge retention and practical skill transfer should be interpreted cautiously when long-term follow-up assessments are not included. Therefore, the incorporation of this tool into teaching processes represents an opportunity to improve technical training, provided that its effects are evaluated through objective and context-specific indicators.

Collectively, the reviewed literature reveals three recurring methodological gaps that limit the applicability of prior findings to real-world automotive maintenance training. First, most AR evaluations have been conducted in university laboratory settings using satisfaction or perceived usability measures, rather than through controlled experimental designs with objectively measured learning outcomes in authentic workplace environments. Second, robustness assessments have been largely restricted to specific hardware configurations — such as head-mounted displays with limited fields of view — without testing marker-based mobile AR systems under the operational conditions of actual maintenance workshops. Third, longitudinal and comparative evidence on training efficiency indicators, such as task completion time and effect size reporting, remains scarce. The present study directly addresses these gaps by implementing a marker-based AR mobile application in a real automotive company, employing a true experimental posttest-only design with random group assignment, and evaluating outcomes through three objective indicators with effect size analysis.

Research Hypotheses. The general hypothesis of this study states that the implementation of a mobile application based on Augmented Reality (AR) improves the training process of automotive maintenance technicians. Furthermore, the following specific hypotheses were formulated: (H1) the AR-based mobile application improves the understanding of technical training content; (H2) the AR-based mobile application reduces the average duration of training sessions; and (H3) the AR-based mobile application increases participant satisfaction with the training method used. These hypotheses were examined through the development and evaluation of an interactive learning tool that allows users to explore diagnostic, repair, and inspection processes in a visual and immersive manner.

II. Theoretical Framework

A. Augmented Reality

Augmented Reality (AR) is a technology that enables the superimposition of digital information—such as three-dimensional models or instructional content—onto the real-world environment through the use of mobile devices or smart displays. In the field of technical education, AR has been incorporated as an innovative tool to simulate complex procedures, visualize internal machine components, and

develop practical skills without directly intervening with physical equipment [12]. This technology enhances the understanding of technical processes, increases technician interaction, and allows for the autonomous repetition of training tasks, resulting in a more dynamic and meaningful learning experience [13].

B. Maintenance in Automotive Systems

Predictive maintenance is based on the continuous monitoring of equipment to identify abnormal conditions that may lead to functional failures. In the automotive sector, this strategy has become a key approach for reducing operational costs, extending component lifespan, and improving vehicle safety. Through the use of smart sensors and machine learning algorithms, it is possible to detect unusual behavioral patterns in engine systems, thereby anticipating potential failures before they occur. When this information is integrated with Augmented Reality (AR) visualizations, technicians can interpret data more quickly and accurately, facilitating decision-making during the maintenance process [13].

C. Mobile Applications

Mobile applications have revolutionized access to knowledge by providing versatile platforms for digital education. In the context of automotive technical training, they enable technicians to interact with simulators, visual manuals, and immersive environments from virtually any location, thereby optimizing autonomous practice. The integration of these applications with Augmented Reality (AR) enables the virtual representation of engines, transmissions, and other vehicle components, promoting both visual and experiential learning. Furthermore, the continuous updates supported by these platforms ensure that training content remains aligned with technological advancements in the sector, thereby fostering more relevant and up-to-date technical training [14].

III. Materials and Methods

A. Type of Research

This study is classified as applied research, as it focuses on analyzing and solving specific problems affecting a particular context. The primary objective was to generate practically useful knowledge aimed at improving processes, optimizing resources, and transforming situations identified as problematic. This approach enabled the integration of theoretical foundations with practical application, directing the results toward the implementation of context-specific solutions. Furthermore, the study aimed to provide scientific evidence to support informed decision-making, the design of improvement strategies, and the innovation of the evaluated procedures, thereby enhancing the relevance and applicability of the findings obtained [15].

B. Research Design

The study employed a true experimental design — specifically, a posttest-only control group design — characterized by the intentional manipulation of an independent variable to observe its effect on a dependent variable under controlled conditions. This design allowed the establishment of cause-and-effect relationships with a higher level of certainty, as it included the random assignment of participants into two groups: an experimental group that received the intervention and a control group that did not receive any treatment. Given that participants were randomly assigned to groups, pre-test measurements were not required, as randomization ensures initial equivalence between the control and experimental groups, thereby controlling for pre-existing differences without the need for baseline measurements. The randomization process helped minimize the influence of extraneous variables and potential biases, thereby ensuring the internal validity of the study. Furthermore, the rigorous control of experimental conditions enabled the comparison of outcomes between both groups, facilitating the identification of effects exclusively attributable to the applied intervention [16].

C. Variables and Operationalization

1) Independent variable: Augmented Reality (AR)

2) Conceptual definition: Augmented Reality (AR) refers to a technology that enables the superimposition of virtual elements onto real-world environments, facilitating the visualization and interaction with three-dimensional content through mobile devices. In educational contexts, it has been established as a tool that enhances comprehension, increases technicians' motivation, and enables the simulation of practical experiences that are difficult to replicate in conventional settings [17].

3) Operational definition: This variable was assessed through the "Implementation of the Mobile Application," assigning the value "NO" prior to the intervention and "YES" after its implementation, using a nominal measurement scale.

4) Dependent variable: Technical Training Performance.

5) Conceptual definition: Technical training performance refers to the level of achievement attained by participants during the training process, determined through the completion of practical activities, evaluation tests, and observed performance outcomes [18].

6) Operational definition: It was measured using three key indicators:

a) Level of understanding of technical procedures,

b) Average time required to complete simulated activities, and

c) Degree of perceived satisfaction with the training process. Data were collected through evaluation records and structured surveys, using ordinal and ratio scales according to the nature of each indicator.

D. Population, Sample, and Sampling

The population is defined as the set of individuals, objects, or events that share common characteristics relevant to the purposes of a research study. It represents the entire group from which information is intended to be obtained and is essential in the study design to determine the scope and representativeness of the results. The characteristics that define a population may include variables such as age, educational level, geographic location, profession, among others. Depending on the methodological approach, research may involve large or specific populations according to the objectives of the study [19].

The population consisted of automotive maintenance technicians working at a company in Trujillo, Peru. The sample comprised a total of 60 technicians selected through purposive sampling, based on criteria such as homogeneity in training level and availability to participate in the proposed activities. Once the sample was defined, participants were randomly assigned to one of two groups of 30 individuals each: an experimental group, which received the AR-based intervention, and a control group, which completed training without the technological tool. This combination of purposive sampling for participant selection and random assignment for group allocation is consistent with true experimental designs in applied research.

E. Data Collection Techniques and Instruments

The research team collected data using quantitative techniques through previously validated instruments. Three main instruments were applied: an evaluation questionnaire to measure content comprehension, a record sheet to calculate the average training time, and a satisfaction survey administered at the end of the process. The collected data were organized and systematized for analysis using descriptive and inferential statistics. Subsequently, Jamovi software was employed to calculate statistical measures such as the mean, standard deviation, and value range. For group comparisons, normality tests were performed, and the Mann–Whitney U test was used due to the non-normal distribution of the collected data. Additionally, effect size was calculated using Rosenthal's r coefficient ($r = Z / \sqrt{N}$) to assess the practical significance of observed differences between groups.

F. Procedures

The study followed several phases based on the Mobile-D methodology, which is considered appropriate for mobile software development projects due to its agile and user-centered approach. In the initial phase, the Augmented Reality (AR)-based mobile application was designed using tools such as Unity 3D, Android Studio, and the Vuforia engine, which enabled the creation of interactive environments and the integration of three-dimensional models. Subsequently, the application was implemented by incorporating functionalities focused on the visualization and manipulation of virtual automotive components. The application was then installed on the technicians' mobile devices, ensuring compatibility with operating systems and proper execution of the programmed functions. Finally, functional and usability testing were conducted to verify application performance and ensure its correct operation within the implementation context.

During the training phase, participants accessed interactive thematic modules that included the visualization of automotive components in 3D. Repetition of activities was permitted as needed, promoting personalized learning. Upon completion of the training process, the corresponding evaluation instruments were administered to measure comprehension levels, time spent, and satisfaction with the training process. Finally, the collected data were processed and analyzed to determine the impact of the technological tool on the training of automotive maintenance technicians.

IV. Software Development Framework

To develop the Augmented Reality (AR)-based mobile application for training automotive maintenance technicians, the team adopted the Mobile-D methodology, an agile approach specifically designed for mobile software projects. This methodology consists of six progressive phases: Exploration, Initialization, Production, Stabilization, System Testing, and Deployment. Each phase addressed a specific set of activities that facilitated the efficient organization of the development process and ensured the adaptation of the product to the needs of end users.

A. Exploration Phase

The Exploration phase marked the starting point for the development of the Augmented Reality (AR)-based mobile application. During this stage, the team analyzed the educational and technological context in which the solution would operate to understand users' needs and their operational environment. The researchers conducted a preliminary needs assessment to identify the limitations of the traditional training model, as well as opportunities for improvement through technological tools. Based on this analysis, the team established the initial system objectives and proposed a mobile-based solution integrating Augmented Reality elements to enhance practical learning.

Additionally, the requirements were identified (see Table I) to guide the software design, distinguishing between those related to the system's core functionalities and those establishing technical conditions aimed at ensuring application performance, usability, and accessibility. These requirements enabled the definition of the main functionalities, technical constraints, and system quality criteria, thereby structuring the development process in an organized manner. Furthermore, the identification of requirements facilitated alignment of the software with end-user needs.

TABLE I. System Requirements

ID	Requirements
R1	The application shall display interactive 3D models of automotive components.
R2	The application shall provide short, dynamic lessons (5–10 minutes) covering specific automotive maintenance concepts.
R3	The application shall enable fast navigation between modules, allowing technicians to access the sections they need to study.
R4	The application shall be user-friendly and intuitive.

B. Initialization Phase

During the Initialization phase, the technical elements that would structure the mobile application were defined, establishing a solid foundation for its subsequent development. Programming tools, development environments, and graphics engines were selected to implement the planned functionalities, ensuring efficient integration of educational content with Augmented Reality (AR) resources. Fig. 1 illustrates the set of components and tools selected for the system's initial architecture.

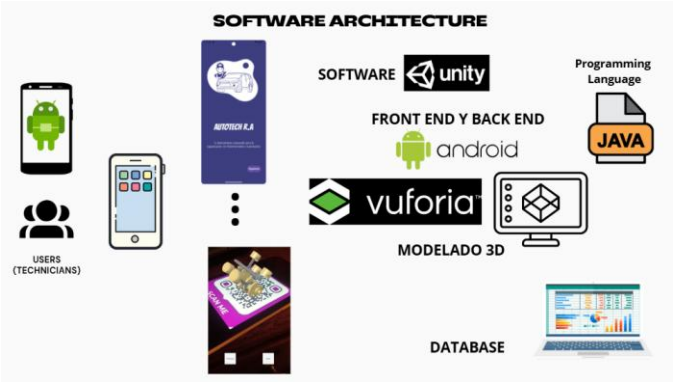


Fig. 1. System architecture.

During the subsequent stage, the overall system architecture (Fig. 1) was structured, considering the different modules that would compose the application. Unity 3D was selected as the primary development environment due to its capability to manage three-dimensional graphical environments and its compatibility with Augmented Reality (AR) tools. For the visualization of augmented objects, the Vuforia engine was integrated, enabling image recognition and the projection of digital models onto real-world markers.

The system logic was developed using C#, a programming language that facilitated dynamic interaction between objects and graphical components. In parallel, Java was utilized to ensure software compatibility with the Android operating system. Additionally, a simple data storage structure was implemented to record user performance information, including assessment results and satisfaction survey responses.

The design also focused on providing intuitive navigation that was accessible and aligned with the needs of the target audience: technical training technicians. A clean visual layout, direct-access buttons, and clear typography adapted to different screen sizes were prioritized. Furthermore, usability and user experience (UX) principles were considered, including the reduction of cognitive load, consistency in graphical elements, and visual feedback in response to user actions.

C. Production Phase

During the Production phase, all components that shaped the functional structure of the Augmented Reality (AR)-based mobile application were developed and integrated. This stage represented a critical milestone, as it enabled the transformation of previously identified technical and pedagogical requirements into interactive and operational modules aimed at enhancing the user's learning experience. The focus of this phase was to combine immersive technology with instructional design principles (see Table II) to build an accessible, functional, and pedagogically effective tool aligned with the needs of technicians in training.

Various thematic modules were designed and implemented, focusing on fundamental components of automotive maintenance, such as engines, valve trains, crankshafts, pistons, suspension systems, among others. Each module was structured to allow users to access essential theoretical content supported by high-quality graphical representations and interactive three-dimensional models.

TABLE II. PRODUCTION PHASE COMPONENTS

ID	Description
Developed Modules	Engines, valve train systems, crankshafts, pistons, and braking systems
Technical Content	Visual and textual guides integrated into each module
Interactive Resources	3D models, animations, navigation buttons, and QR codes
Functional Logic	Modular flow organized by thematic levels, incorporating touch-based and visual interaction
Technical Validation	Review conducted by automotive industry professionals
AR Integration	Linking 3D models to visual markers for augmented visualization

The development of these modules followed a user-centered approach. The development team built the application to be intuitive, clear, and fluid, enabling participants to move easily between content sections. The main interface of the application, shown in Fig. 2 and 3, presents an organized view of the available modules, where each option is accompanied by a representative image and a direct-access button. The interface is presented in Spanish, as the application was developed for a Spanish-speaking technical education context. The visual layout allows users to quickly identify topics, promoting autonomous access to the learning materials.

Upon entering each module, as shown in Fig. 4, users were presented with a general introduction to the topic, followed by a detailed description of the system components, their operation, and maintenance procedures. This information was delivered through simplified technical text, designed to be reinforced through the exploration of Augmented Reality (AR) models, which responded to interaction commands such as rotation, zooming, and perspective changes.

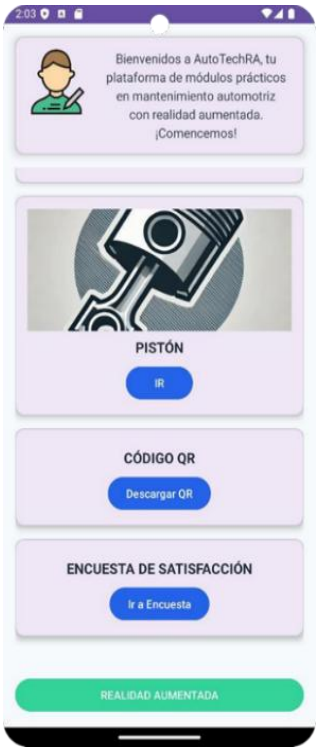


Fig. 2. Main menu interface.

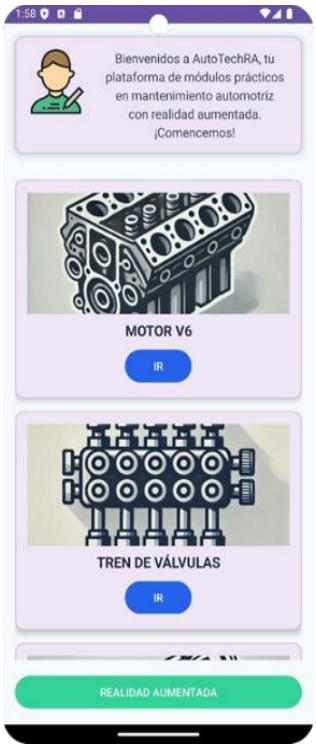


Fig. 3. Main menu interface layout.

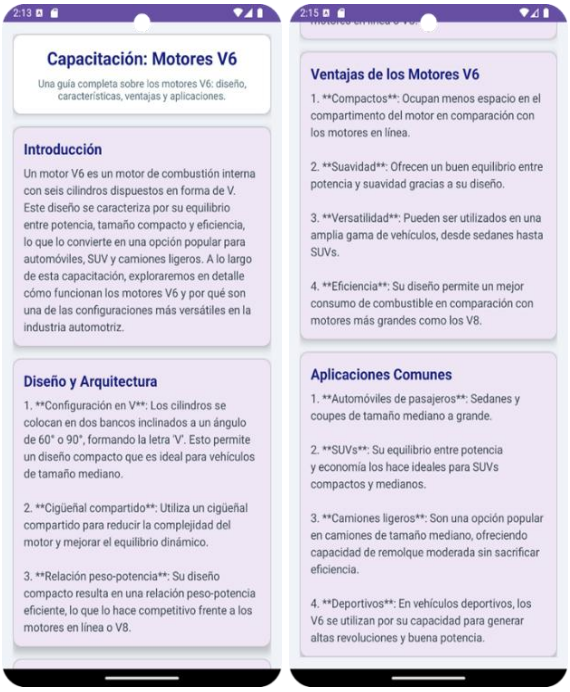


Fig. 4. Module information interface.

From a pedagogical perspective, the design prioritized visual clarity, logical content progression, and the sequential execution of learning activities. Animations and intuitive navigation controls were incorporated to enable smooth transitions between modules, thereby reinforcing learning continuity. Additionally, complementary resources were integrated to enrich the in-app navigation experience and strengthen the connection between visual and conceptual content.

Among the implemented resources, navigational buttons (Fig. 5 and 6) were strategically positioned to facilitate movement between sections, allowing users to advance, return, or restart each module without difficulty. This functionality was specifically designed to ensure user autonomy during the learning process, reducing the need for external assistance and promoting independent exploration of the content.



Fig. 5. AR-based 3D visualization of engine components.



Fig. 6. AR-based 3D visualization of brake system components.

Finally, the inclusion of a QR code (Fig. 7) aimed to promote learner autonomy by enabling technicians to review content at different times and across various devices. This functionality proved particularly useful for reinforcing key concepts, visualizing procedures from different perspectives, and accessing reliable resources through a single scan using the mobile device camera.

Continuous validation was conducted with automotive industry specialists, who reviewed the technical accuracy of the models and the content presented in each module. Adjustments were also made to the user interface (UI) and user experience (UX) components to ensure coherent, accessible navigation aligned with the characteristics of the target audience. In summary, the Production phase resulted in the consolidation of an operational version of the application, composed of complete thematic modules, integrated Augmented Reality (AR) functionalities, and a user-centered design.



Fig. 7. User interface for content access.

D. Stabilization Phase

The Stabilization phase aimed to ensure that the developed mobile application operated correctly, stably, and efficiently before being deployed to end users. During this stage, the team conducted rigorous internal testing to identify technical errors, verify system performance, and optimize the user experience under different operational conditions. The evaluation focused on aspects such as 3D model loading time, stability of core functionalities, navigation fluidity between modules, and system responsiveness across different mobile devices.

These tests enabled the identification of necessary adjustments, including graphic resource compression, animation optimization, and code debugging to prevent potential failures or performance slowdowns during execution. The application's behavior when interacting with Augmented Reality (AR) elements was also evaluated. Performance and robustness testing was conducted on Android and iOS mobile devices, recording an average initial loading time of 3.5 seconds, RAM usage of approximately 250 MB, and AR interaction response times under 2 seconds. QR marker recognition and 3D model projection were evaluated under the indoor lighting conditions typical of an automotive workshop environment, confirming stable tracking performance across the tested devices.

Finally, the phase concluded with an optimized application ready for evaluation in a real-world environment. The actions undertaken during this stage were essential to prevent errors during field use and to ensure a positive educational experience based on mobile technology and Augmented Reality (AR). Additionally, preliminary functional tests were conducted to verify system stability, correct visualization of virtual objects, and appropriate interaction with AR markers.

V. Results

A total of 60 automotive maintenance technicians participated in this study, with 30 assigned to the control group and 30 to the experimental group.

A. Level of Comprehension of Technical Content (NCC)

The assessment of the level of comprehension of technical content revealed marked differences between the two groups. In the control group, participants were predominantly categorized at the "Beginning" and "In Progress" performance levels. Conversely, the experimental group showed a higher concentration of participants at the "Expected Achievement" and "Outstanding Achievement" levels. The Shapiro–Wilk normality test indicated that the data did not follow a normal distribution in either group (Control Group: $p = 0.024$; Experimental Group: $p = 0.029$). Due to this non-normal distribution, the non-parametric Mann–Whitney U test was employed for group comparison. The test yielded a p-value of 0.001 (at a 95% confidence level), indicating a statistically significant difference between the control and experimental groups regarding the level of comprehension of technical content. Furthermore, the effect size calculated using Rosenthal's r coefficient yielded $r = 0.86$, indicating a large practical effect and confirming that the observed differences were not only statistically significant but also educationally meaningful.

B. Average Training Time (ATT)

The Average Training Time (ATT) differed substantially between groups. The control group technicians required an average of 10,862 seconds (approximately 181 minutes) to complete all training activities. The experimental group technicians completed the same activities in an average of 5,784 seconds (approximately 97 minutes), representing a reduction of 5,078 seconds, or 46.6% less time than the control group. Shapiro–Wilk testing revealed non-normal data distribution (Control Group: $p < 0.001$; Experimental Group: $p = 0.054$). The Mann–Whitney U test comparing both groups of participants yielded a p-value of 0.001, establishing statistical significance at the 95% confidence level. This reduction in training duration occurred while maintaining the same training objectives and assessment criteria across both groups. The effect size analysis yielded $r = 0.87$, indicating a large practical effect and confirming that the reduction in training time represented a substantial and meaningful improvement beyond statistical significance (Table III).

TABLE III. Summary of Statistical Results for Primary Indicators

Metric / Group	NCC[1]	ATT[2]	NCS[3]
Control Group	Beginning / In Progress	10,862 s (≈)	M = 2.6 (5-point scale)
Experimental Group	Expected Achievement / Outstanding Achievement	5,784 s(≈97min)	M = 4.0 (5-point scale)
p-calue	P = 0.001	P = 0.001	P = 0.001
Effect Size (Rosenthal's r)	0.86 (large)	0.87 (large)	0.86 (large)
Result	Significant	Significant	Significant

C. Participant Satisfaction Level (NSC)

Satisfaction was measured on a 5-point Likert scale (1 = Very Dissatisfied; 5 = Very Satisfied). The control group technicians achieved a mean satisfaction score of 2.6, while the experimental group technicians obtained a mean score of 4.0. This 1.4-point difference represents a 54% increase in satisfaction levels in the experimental group compared to the control group. The Mann–Whitney U test comparing satisfaction levels between participants yielded a p-value of

0.001 (at a 95% confidence level), confirming a statistically significant difference in favor of the experimental group technicians.

The effect size calculated using Rosenthal's r coefficient yielded r = 0.86, indicating a large practical effect and demonstrating that the improvement in satisfaction levels was not only statistically significant but also of considerable practical relevance.

VI. Discussion

As detailed in the Results section, the experimental group achieved significantly higher scores in the comprehension of technical content (NCC) and participant satisfaction (NSC), while reducing the average training time (ATT) by 46.6% (p = 0.001). Regarding the first indicator, the level of comprehension of technical content (NCC) showed a marked improvement in the experimental group that utilized the Augmented Reality (AR) application. In contrast, the control group remained predominantly at performance levels categorized as "Beginning" and "In Progress," whereas the experimental group reached higher levels such as "Expected Achievement" and "Outstanding Achievement," confirming a statistically significant difference between groups. The large effect size obtained (r = 0.86) further supports the educational relevance of these differences, indicating that the AR-based intervention produced a substantial improvement beyond mere statistical significance. These findings highlight that the immersive and interactive nature of AR-based tools enhanced the assimilation of complex technical content, allowing technicians to develop a deeper conceptual understanding than that achieved through conventional training methods.

These results are consistent with prior studies, such as Aoufy Khadija et al. [20], which demonstrated that participants trained using immersive technologies—like AR and Virtual Reality (VR)—achieved significantly higher post-training assessment scores. This evidence reinforced the positive impact of immersive technologies on knowledge retention, comprehension, and practical application.

Moreover, within the context of this study, the incorporation of interactive and immersive environments promoted active technician participation, increased motivation for learning, and facilitated the visualization of complex concepts that are difficult to grasp through traditional methods. These findings suggest that integrating AR-based tools may represent a promising pedagogical strategy to improve technical training outcomes, although further research is needed to confirm these effects in broader educational contexts.

Concerning the second indicator, Average Training Time (ATT), the experimental group demonstrated a substantial improvement in efficiency, completing the training activities in significantly less time than the control group. This improvement suggested that the structured, module-based design of the AR application—which allowed technicians to interact directly with three-dimensional representations of automotive components—reduced the cognitive effort required to understand procedural sequences, thereby accelerating task completion without compromising learning quality. The effect size analysis (r = 0.87) confirmed that this reduction was not only statistically significant but also of large practical magnitude, reinforcing the efficiency gains attributed to the AR-based intervention.

These findings aligned with the study conducted by Yıldırım and Yıldırım [8], which reported that AR and VR technologies reduced the time needed to complete complex technical tasks by 10% to 15%, attributed to direct and detailed visualization of processes, three-dimensional component representation, and the superimposition of digital information onto the physical environment.

Additionally, within this study, immersive tools enabled participants to interact with virtual models in real time, strengthening procedural and cognitive skill acquisition. These results suggest that AR-based applications may contribute to improving training efficiency, though generalization to broader professional settings requires further investigation.

Finally, regarding technician satisfaction (NSC), a clear difference emerged between groups. While the control group reported predominantly low satisfaction scores, the experimental group achieved a mean score of 4.0 on a 5-point Likert scale, reflecting a statistically significant improvement (p = 0.001). The large effect size (r = 0.86) confirmed that this improvement was of considerable practical relevance, not merely a statistically detectable difference. This outcome demonstrated that the immersive and autonomous nature of AR-based training positively influenced user perception, fostering engagement and motivation. Previous studies, such as Ojer et al. [21], supported these results, showing that immersive environments increased perceived usability and motivation, with satisfaction levels reaching up to 90% compared to approximately 65% for traditional teaching methods.

Interaction with virtual and augmented environments also enhanced engagement, sustained attention, and positive perceptions of the learning process, which translated into a greater willingness to acquire technical knowledge and develop practical skills. Consequently, the integration of immersive technologies in technical training programs represented an innovative strategy with strong potential to optimize learning effectiveness and enhance user experience.

Overall, the implementation of the AR-based mobile application produced statistically significant improvements across all three evaluated indicators: comprehension of technical content, training efficiency—with a 46.6% reduction in average time—and participant satisfaction (p = 0.001 for all cases). These outcomes suggest that, within the context of this study, AR-based tools not only optimized instructional processes but also enhanced motivation and engagement among technicians, indicating a positive influence on learning quality and willingness to acquire new technical competencies.

The incorporation of immersive technologies promoted an interactive and user-centered learning experience within the evaluated training context. These findings suggest that AR-based mobile applications may contribute to aligning technical training with contemporary industry demands, though broader conclusions regarding industry-wide impact would require

studies conducted across more diverse organizational and regional settings.

Despite these promising findings, the study presented several limitations. First, the sample was drawn from a single automotive company in Trujillo, Peru, which may limit the generalizability of the results to technicians working in different organizational contexts, training environments, or regions. Second, the study did not assess long-term knowledge retention, leaving uncertain whether the observed learning gains persisted over time. Third, outcomes may have been partially influenced by novelty effects associated with the introduction of new technology, potentially inflating initial motivation and satisfaction scores. Fourth, although the application was tested across multiple device models, robustness testing was limited to indoor workshop lighting conditions, without evaluating marker-based AR performance under varying illumination or outdoor environments. Future research should address these limitations by incorporating larger and more diverse samples across multiple organizations, longitudinal follow-up assessments, control conditions that account for novelty effects, and robustness testing across diverse environmental settings.

VII. Conclusion

Regarding the level of technical content comprehension, it was concluded that the use of the Augmented Reality (AR)-based mobile application significantly enhanced learning outcomes among automotive maintenance technicians. The group that utilized the application predominantly achieved performance levels categorized as “Expected Achievement” and “Outstanding Achievement,” in contrast to the group that did not use the tool. This difference was statistically supported by the Mann–Whitney U test ($p = 0.001$), allowing acceptance of the alternative hypothesis at a 95% confidence level and suggesting that, within the context of this single-company study, the implemented technology had a positive impact on the assimilation of complex technical content, though these findings should be interpreted with caution given the limited deployment scope.

Regarding the average training time, a notable improvement in efficiency was observed. The experimental group completed the training sessions in an average of 5,784 seconds, whereas the control group recorded 10,862 seconds. This 46.6% reduction in training time suggests that the application optimized the instructional process without compromising learning quality under the conditions of this study, though longitudinal validation would be needed to confirm sustained efficiency gains. The statistical significance of this difference was validated through the Mann–Whitney U test ($p = 0.001$), providing statistical evidence that the technological implementation reduced training duration within the evaluated context.

Regarding participant satisfaction, a considerable increase was observed. The group that used the application reported higher satisfaction levels, achieving a mean score of 4.0 on a 5-point scale, whereas the control group obtained an average of 2.6 points. These results reflect not only improved reception of the training process but also greater motivation and engagement among the technicians. As in the previous indicators, the difference was statistically significant ($p = 0.001$), reinforcing the conclusion that, within this study's context, the AR-based educational approach was positively received by participants, though broader adoption would require validation across more diverse training environments and longer evaluation periods.

References

[1] Arena, F., Collotta, M., Luca, L., Ruggieri, M., & Termine, F. G. (2022). Predictive Maintenance in the Automotive Sector: A Literature Review. Mathematical and Computational Applications, 27(1), 2. https://doi.org/10.3390/mca27010002

[2] Cordero-Guridi, J.-d.-J., Cuautle-Gutiérrez, L., Alvarez-Tamayo, R.-I., & Caballero-Morales, S.-O. (2022). Design and development of a I4.0 engineering education laboratory with virtual and digital technologies based on ISO/IEC TR 23842-1 standard guidelines. Applied Sciences, 12(12), 5993. https://doi.org/10.3390/app12125993

[3] Boboc, R. G., Gîrbacia, F., & Butilă, E. V. (2020). The Application of Augmented Reality in the Automotive Industry: A Systematic Literature Review. Applied Sciences, 10(12), 4259. https://doi.org/10.3390/app10124259

[4] Liu, X.-W., Li, C.-Y., Dang, S., Wang, W., Qu, J., Chen, T., & Wang, Q.-L. (2022). Research on Training Effectiveness of Professional Maintenance Personnel Based on Virtual Reality and Augmented Reality Technology. Sustainability, 14(21), 14351. https://doi.org/10.3390/su142114351

[5] Palmarini, R., Erkoyuncu, J. A., Roy, R., & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. Robotics and Computer-Integrated Manufacturing, 49, 215–228. https://doi.org/10.1016/j.rcim.2017.06.002

[6] Pech, M., Vrchota, J., & Bednář, J. (2021). Predictive Maintenance and Intelligent Sensors in Smart Factory: Review. Sensors, 21(4), 1470. https://doi.org/10.3390/s21041470

[7] Husár, J., Knapčíková, L. Implementation of Augmented Reality in Smart Engineering Manufacturing: Literature Review. *Mobile Netw Appl 29, 119–132* (2024). https://doi.org/10.1007/s11036-023-02121-x

[8] Yıldırım, B., & Yıldırım, S. (2024). Concepts, innovative technologies, learning approaches and trend topics in Education 4.0: A scoping literature review. Social Sciences & Humanities Open, 9, 100902. https://doi.org/10.1016/j.ssaho.2024.100902

[9] Zhang, Y. (2024). Research on vocational education practice teaching mode under the background of Internet Plus. International Journal of Web-Based Learning and Teaching Technologies, 20(1), 1–18. https://doi.org/10.4018/IJWLTT.382589

[10] Yaroson, E. V., Chowdhury, S., Mangla, S. K., et al. (2024). Unearthing the interplay between organisational resources, knowledge and industry 4.0 analytical decision support tools to achieve sustainability and supply chain wellbeing. Annals of Operations Research, 342, 1321–1368. https://doi.org/10.1007/s10479-024-05845-5

[11] Fombona, J., Pascual-Sevillano, M., & González-Videgaray, M. (2017). M-learning and augmented reality: A review of the scientific literature on the WoS repository. Comunicar, 52, 63–72. https://doi.org/10.3916/C52-2017-06

[12] Ghobadi, M., Shirowzhan, S., Ghiai, M. M., Mohammad Ebrahimzadeh, F., & Tahmasebinia, F. (2023). Augmented Reality Applications in Education and Examining Key Factors Affecting the Users’ Behaviors. Education Sciences, 13(1), 10. https://doi.org/10.3390/educsci13010010

[13] Ahmad, A. R., Ali, D. F., Othman, N. F., et al. (2024). Technological innovations in welding education: A review of teaching and learning approaches. International Journal of Academic Research in Progressive Education and Development, 13(4). http://dx.doi.org/10.6007/IJARPED/v13-i4/23801

[14] Chiang, F. K., Shang, X., & Qiao, L. (2022). Augmented reality in vocational training: A systematic review of research and applications. Computers in Human Behavior, 129, 107125. https://doi.org/10.1016/j.chb.2021.107125

[15] Pineda, E. B., de Alvarado, E. L., & de Canales, F. H. (2020). Metodología de la investigación: Manual para el desarrollo de personal de salud (2.ª ed.). Organización Panamericana de la Salud. https://iris.paho.org/handle/10665.2/3132

[16] Ramos-Galarza, C. (2021). Editorial: Diseños de investigación experimental. CienciAmérica, 10(1), 5–6. https://doi.org/10.33210/ca.v10i1.356

[17] Kamińska, D., Zwoliński, G., Laska-Leśniewicz, A., Raposo, R., Vairinhos, M., Pereira, E., Urem, F., Ljubić Hinić, M., Haamer, R. E., & Anbarjafari, G. (2023). Augmented Reality: Current and New Trends in Education. Electronics, 12(16), 3531. https://doi.org/10.3390/electronics12163531

[18] Tessaro, I., Mariani, V. C., & Coelho, L. d. S. (2020). Machine Learning Models Applied to Predictive Maintenance in Automotive Engine Components. Proceedings, 64(1), 26. https://doi.org/10.3390/IeCAT2020-08508

[19] Arrogante, O. (2022). Técnicas de muestreo y cálculo del tamaño muestral: Cómo y cuántos participantes debo seleccionar para mi investigación. Enfermería Intensiva, 33(1), 44–47. https://doi.org/10.1016/j.enfi.2021.03.004

[20] Aoufy, K., Amato, C., Balestri, C., & Longobucco, Y. (2023). Virtual Reality and Augmented Reality Training in Disaster Medicine Courses for Students in Nursing: A Scoping Review of Adoptable. Tools. Behavioral Sciences, 13(7), 616. https://doi.org/10.3390/bs13070616

[21] Ojer, J., Alvarez, H., Serrano, I., Saiz, F. A., Barandiaran, I., Aguinaga, D., Querejeta, L., & Alejandro, D. (2020). Projection-Based Augmented Reality Assistance for Manual Electronic Component Assembly Processes. Applied Sciences, 10(3), 796. https://doi.org/10.3390/app10030796

CFD-Based Aerothermal Design Verification of Coffee Cabinet Dryer

Gerry M. Castillo[1], Aeron R. Mojica[2], Jhon Heron P. Carsocho[3],
Ray Rojan C. Romeroso[4], Sheryl Dinglasan-Fenol[5], Al Eugene L. Torres[6], Gee Jay C. Bartolome[7]
College of Engineering and Information Technology, Cavite State University, Indang, Cavite, Philippines[1]
National Coffee Research, Development and Extension Center, Cavite State University, Indang, Cavite, Philippines[2, 3, 4, 6]
Research Center, Cavite State University, Indang, Cavite, Philippines[5]
Research and Extension, Cavite State University – CCAT, Rosario, Cavite, Philippines[7]

***Abstract*—Traditional sun drying remains the primary practice among smallholder coffee farmers in the Philippines, often leading to an inefficient and inconsistent drying process. While transitioning to mechanical drying improves product quality, achieving uniform airflow in cabinets with multiple trays often requires costly post-fabrication modifications, resource-intensive trials, and repetitive field testing to fine tune the performance of the system. To reduce inefficient and repetitive trial-and-error procedures, a pre-fabrication design verification of a small-scale mechanical cabinet dryer was conducted using COMSOL Multiphysics. This study explores how the dryer's geometry influences its aerodynamic and thermal performance under baseline, no-load conditions prior to fabrication. The system was simulated to deliver an inlet air velocity of 1.56 m/s and a controlled heat flux to maintain 40°C. Results revealed steady pressure equalization within the plenum chamber, which successfully mitigated non-uniform airflow. As a result, a relatively high Velocity Uniformity Index of 0.84 was achieved across the tray layers. In addition, internal turbulent mixing contributed to a more stable thermal profile, reaching a near-ideal Temperature Uniformity Index of 0.99. These findings reveal that the core geometry of the dryer performs well from an aerodynamic standpoint. By achieving flow equalization in a no-load setting, the study provides an optimized baseline design, helping reduce the need for extensive trial-and-error during future prototyping and testing under actual operating conditions.**

***Keywords*—COMSOL Multiphysics; coffee dryer; heat transfer; design verification; aerothermal simulation; performance characterization**

I. Introduction

Drying is considered the most common and oldest technique for preserving agricultural crops, where moisture content and water activity are reduced to minimize chemical, biochemical, and microbiological deterioration [1][2][3]. This postharvest process is recognized as the most critical step affecting the final quality of the coffee [4][5][6][7]. Apart from influencing the physical appearance of the coffee beans, drying can also influence the quality of the end-product of the hulling process [5][8].

Open sun drying is still widely used nowadays for drying agricultural products even in various developing countries with the goal of lowering the moisture content of the beans [9][10][11]. However, this traditional approach poses several risks including contamination, longer drying period and unpredictable weather patterns and conditions [11]. Despite its low cost and simplicity that attracts coffee growers [11], sun drying approach is not recommended due to numerous factors that could threaten the quality of the product [10].

Sun drying, as a weather-dependent process, could take up two to three weeks which makes it a slow process that requires regular stirring to ensure even drying [11]. Drying coffee on open pavements usually takes several days, depending on the type of processing. Washed coffee generally dries in about six to seven days, pulped natural or semi-washed coffee needs roughly eight to nine days, and natural or dry-processed coffee requires the longest, often around 12 to 14 days [12]. Apart from that, the harvesting season for coffee coincides with the rainy season of the year. This makes the process labor-intensive, as the commodity should be covered during adverse weather conditions [11]. Furthermore, extended drying periods increase the likelihood of microorganisms, molds, and mycotoxins developing, which can compromise the quality of the product [10]. This includes contamination by dust, insects, and other domestic animals. All these factors prevent the coffee farmers from utilizing alternative, and more efficient drying methods when traditional sun drying is not possible [12].

Aside from traditional open sun drying, some common methods for drying the coffee cherries are the solar and mechanical drying [10][13][14][15]. The use of solar energy in drying later expanded to include solar dryers. These systems improve drying efficiency and help reduce problems such as mold growth and microbial contamination. However, because they rely entirely on sunlight, unstable weather can still hinder their performance. In coffee processing, solar drying offers an advantage: temperatures generally stay below 50°C, which preserves the embryo inside the bean. Once this temperature is exceeded, the embryo can be damaged, leading to rapid internal deterioration and lower overall quality [10][16]. On the other hand, mechanical drying serves to accelerate the slowest segment of the drying curve, specifically the drop from 15% to 11% moisture, thereby minimizing the risk of unnecessary fermentation [12]. Moreover, mechanical dryers were found to preserve significantly higher levels of phenolic compounds and antioxidant activity compared to sun drying [17]. Through the use of a mechanical dryer, water activity was maintained below 0.6 and with a stable moisture content of 7.7–10.4%, conditions that effectively inhibit microbial growth and prevent mycotoxin formation [18].

These advantages, however, are highly dependent on the internal geometry of the cabinet and its capacity to deliver uniform airflow and stable temperatures across all tray levels. Non-uniform drying in such dryers is primarily caused by poor airflow distribution through the chamber, which can be improved by optimizing the flow path and ensuring more uniform velocity across trays [19]. Repetitive field testing through trial-and-error is costly and time-consuming, making simulation a practical alternative for evaluating dryer performance before physical prototypes are developed.

This study computationally characterizes the aerodynamic and thermal performance of a small-scale mechanical coffee cabinet dryer using COMSOL Multiphysics under steady-state, no-load conditions. Isolating the pressure and velocity distributions across the plenum and intermediate ducts enables a direct evaluation of the system's capacity to achieve foundational flow equalization. Verifying that the core geometry performs well from an aerodynamic standpoint validates the underlying structural design, helping reduce the need for extensive trial-and-error during future physical prototyping and testing under actual operating conditions.

II. Literature Review

Before the physical fabrication, computational simulation enables a controlled environment to study and analyze airflow behavior, heat distribution, and other key performance parameters while minimizing uncertainties in the design process. Understanding and optimizing these dynamics is crucial for producing high-quality output, extending shelf life, and supporting broader goals of food safety and sustainability [17]. This goal can be achieved through proper modelling of heat and airflow dynamics across the system [20].

In recent years, Computational Fluid Dynamics (CFD) has been recognized as a critical tool for diagnosing and correcting non-uniform airflow, which is the primary cause of uneven drying in cabinet systems, and recent literature strengthens its role in dryer optimization. Advancements in CFD modeling have shifted attention toward achieving uniform velocity distribution, with iterative design refinements based on simulation outputs becoming standard practice for improving dryer performance [21]. Flow stratification and non-uniform drying are well-documented drawbacks of cabinet drying systems. However, optimizing internal geometry through the precise modeling of heat and airflow dynamics can significantly mitigate these issues. Existing studies show that optimizing cabinet geometry and airflow pathways not only enhances uniformity but can generate substantial efficiency gains, including up to a 50 percent reduction in drying time and overall energy use, especially when combined with heat pump technology [19][21][22][23][24][25]. Furthermore, the reliability of this computational approach is confirmed by experimental validations, which consistently demonstrate less than a 10 percent relative error between simulated and physical values when predicting complex airflow behaviors and recirculation zones.

For instance, a CFD-integrated sequential linear programming framework was utilized to systematically evaluate the influence of supply fan positions on airflow uniformity in a multi-tray cabinet dryer [22]. The study achieved a highly uniform airflow range of 0.82–1.9 m/s, underscoring the effectiveness of CFD-based geometric optimization. However, it was emphasized that the absence of thermal coupling and experimental validation in purely aerodynamic simulations is a crucial limitation since temperature variations may become the next limiting factor. Therefore, comprehensive optimization frameworks must integrate both aerodynamic and thermal uniformity.

To establish a reliable aerothermal baseline before introducing the complex variables of agricultural products, several studies evaluated drying systems under no-load, steady-state conditions. A generalized methodology for the thermal testing of cabinet and solar dryers in no-load conditions has been developed, utilizing a no-load performance index (NLPI) to accurately compare structural designs [25]. Building on this principle, explicitly evaluating the temperature and air-velocity distribution inside an empty drying chamber has been demonstrated as a necessary prerequisite to verify a system's thermal and aerodynamic behavior before loading the product [24]. Furthermore, computational tools such as COMSOL Multiphysics have been successfully utilized for the finite-element modeling of dryers under no-load conditions, with simulated thermal data validated against physical experimental tests [23]. Overall, these insights reveal that CFD functions both as a diagnostic framework and a design optimization tool, enabling evidence-based improvements in drying uniformity, system efficiency, and the overall performance of cabinet dryers.

III. Materials and Methods

A. Computational Domain and Dryer Geometry

The coffee cabinet dryer was developed using Autodesk Fusion 360 computer-aided design (CAD) software. The 3D computational domain encompasses the internal fluid volume of the system's major assemblies: the plenum chamber, air routing duct, and the double-walled, fiberglass-insulated drying chamber containing a 5-layer stainless-steel tray assembly (Fig. 1).



Fig. 1. Full 3D CAD model of the proposed mechanical coffee cabinet dryer.

To establish a reliable baseline before introducing the complex variables of agricultural products, the internal domain was modeled under no-load conditions. The tray levels were treated as unobstructed fluid volumes, isolating the aerodynamic performance of the structural geometry to ensure that any observed flow stratification is a direct result of the cabinet's

architecture rather than the variable hydraulic resistance introduced by coffee beans (Fig. 2).

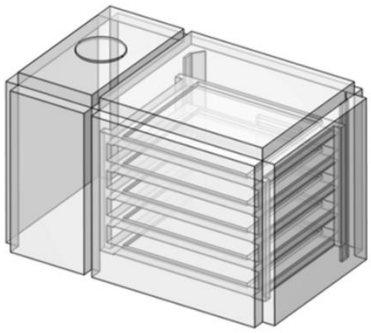


Fig. 2. Extracted computational domain representing the internal fluid volume for CFD analysis.

B. Design Assumptions and Governing Equations

The numerical simulation was executed in COMSOL Multiphysics based on several simplifying physical assumptions to ensure computational efficiency while maintaining acceptable aerothermal accuracy:

1) Incompressible flow: Airflow operates at low Mach numbers; therefore, density variations are negligible, and the fluid is treated as incompressible.

2) No internal heat generation: Heating originates externally from the simulated heating element; there is no volumetric heat generation within the drying chamber itself (Q = 0).

3) Adiabatic boundaries: Because the physical prototype utilizes thick fiberglass lagging, the external walls are treated as perfectly insulated.

The physical phenomena inside the dryer were governed by the conservation of mass, momentum, and energy. Mass conservation for the incompressible fluid is defined by the continuity equation:

$$\nabla \cdot u = 0 \tag{1}$$

Because the forced convection driven by the supply fan induces high-velocity gradients and internal recirculation, the airflow was characterized as turbulent. Momentum transport was resolved using the Reynolds-Averaged Navier-Stokes (RANS) formulation:

$$\rho\left(\frac{\partial t}{\partial u} + u \cdot \nabla u\right) = -\nabla p + \nabla \cdot [(\mu + \mu t)(\nabla u + (\nabla u)T)] \tag{2}$$

Turbulence effects were modeled using the Realizable k–ε model, which provides robust predictions for internal flows with rotational features and flow separation. The turbulent eddy viscosity (μ_t) is defined as:

$$\mu t = C\mu\frac{k^2}{\varepsilon} \tag{3}$$

To fully describe the thermal behavior inside the dryer, the general convection-conduction energy balance was applied to couple the thermal distribution with the velocity field:

$$\rho C p u \cdot \nabla T + \nabla \cdot q = Q + Qp + Qvd \tag{4}$$

where, $\rho C_p u \cdot \nabla T$ represents the dominant convective heat transport by the airflow, and $\nabla \cdot q$ models the conductive heat transfer through the solid boundaries of the system. The parameter Q corresponds to the external volumetric heat source, whereas Q_p represents the work associated with pressure changes, and Q_{vd} accounts for heating due to viscous dissipation.

Under the low Mach number and low-velocity operating conditions considered in this study, the effects of pressure work (Q_p) and viscous shear heating (Q_{vd}) are negligibly small. Consequently, these terms are systematically treated as zero within the numerical solver.

C. Boundary Conditions and Thermal Control Logic

The simulation boundary conditions were established to precisely replicate the intended operational parameters of the physical prototype. Aerodynamically, a velocity inlet boundary condition of 1.56 m/s was applied at the air duct entry, corresponding to a steady volumetric flow rate of 0.03 m³/s. The exhaust vent at the top of the cabinet was modeled as a pressure outlet set to standard atmospheric pressure, expressed as 0 Pa gauge pressure, permitting the free discharge of the circulating air into the ambient surroundings. All internal structural walls, baffles, and tray supports were assigned a standard no-slip condition (u = 0).

To establish the internal thermal environment, a fixed air temperature boundary condition of 373.15 K (100°C) was imposed at the outlet of the heating plenum chamber, immediately upstream of the connecting air duct leading into the drying chamber. This represents the temperature of the conditioned process air supplied by the heating system prior to distribution within the cabinet. Because the physical cabinet utilizes thick fiberglass lagging, the external boundaries were approximated as adiabatic to isolate the internal thermal distribution. However, to account for minor structural conductive losses and accurately model the target internal drying environment of 313.15 K (40°C), a constant surface heat flux of 180 W/m² was applied to the external walls.

Furthermore, to evaluate the transient thermal response of the system under operational conditions, the heating element was governed by an ON/OFF thermostat rule with hysteresis. This control logic dynamically regulated the average domain temperature (T_{avg}) by defining the volumetric heat source density, $q''(t)$ in W/m³, based on predefined lower (T_{on}) and upper (T_{off}) temperature thresholds. The governing control logic for the heat source is mathematically expressed as a piecewise function:

$$q''(t) = \begin{cases} q_0 & if\ T_{avg} \leq T_{on} \\ 0 & if\ T_{avg} \geq T_{off} \\ Interpolated & if\ T_{on} < T_{avg} < T_{off} \end{cases} \tag{5}$$

where, T_{on} is the lower temperature threshold, T_{off} is the upper temperature threshold, and q_0 is the maximum heat source capacity. This hysteresis formulation ensures realistic thermal cycling, preventing high-frequency artificial oscillation in the

numerical solver while accurately reflecting the thermal buffer capacity of the physical drying chamber.

D. Mesh Generation and Grid Independence

The resolution of the computational model, determined by the mesh parameters, governs how the Finite Element Method breaks down the drying chamber geometry into simple elements. Achieving greater accuracy requires a finer mesh, but this refinement directly increases both the necessary computation time and memory resources [20]. The simulation pipeline used in this study was adapted from the study of [20], which involves the CFD model analysis of a multipurpose dryer (Fig. 3).

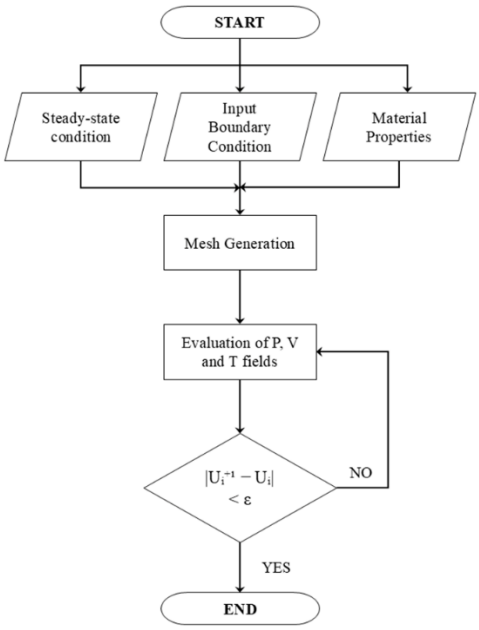


Fig. 3. Adapted simulation pipeline based on the framework presented in [20].

With the governing physics and operational boundaries fully defined, the computational domain required spatial discretization to resolve the flow and temperature fields. A physics-controlled, unstructured mesh utilizing 3D tetrahedral domain elements and triangular boundary elements was generated across the three primary domains elaborated in Table I.

TABLE I. MESH STATISTICS FOR THE THREE MAJOR DOMAINS

Computational Component	Domain Elements (Tetrahedra)	Boundary Elements (Triangles)	Edge Elements (Edges)	Total Elements
Drying Chamber	417,186	46,268	3,114	466,599
Plenum Chamber	27,492	7,642	800	35,394
Air Duct	70,820	9,920	680	81,420

To guarantee that the predicted aerothermal fields were independent of discretization error, a grid independence study was performed. Four progressively refined meshes were evaluated, consisting of approximately 495,566; 527,053; 606,877; and 749,110 total elements (Fig. 4).

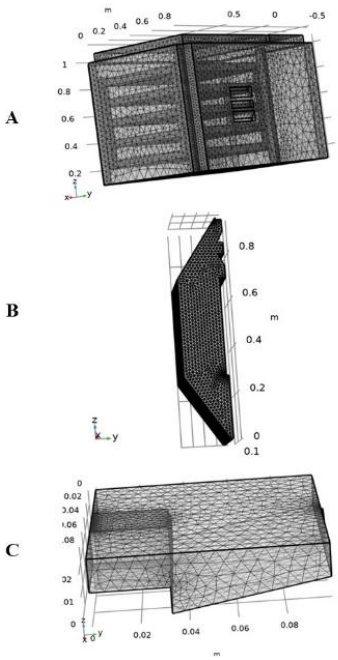


Fig. 4. Generated mesh for each computational domain: (A) Drying chamber; (B) Air duct; and (C) Plenum chamber.

The placement of virtual sensors was strategically designed to capture the full gradient of the aerothermal environment. Specifically, point probes were positioned on all interior boundary surfaces of the drying chamber's domain to monitor wall-proximate effects. Furthermore, probes were localized at the geometric center of each of the tray layers, ensuring that the refinement analysis accounted for the critical airflow and temperature variations within the product-loading zones.

To assess convergence across these iterations, six sensitive virtual probe locations were defined within the computational domain including the heated air inlet, the front and center of the third tray, and representative chamber walls to systematically monitor temperature variations across each successive refinement.

E. Simulation Strategy

The investigation was divided into two distinct computational phases. A steady-state simulation characterized the baseline aerodynamic behavior of the system, with emphasis on airflow distribution, static pressure equalization, and overall flow uniformity. A subsequent transient simulation spanning 60 minutes evaluated the system's thermal response and the performance of the thermostat control logic.

The segregated solver employed pseudo-time stepping with Anderson acceleration, successfully achieving stable convergence for both fluid flow and turbulence variables. To capture the transient thermal response during this 60-minute evaluation, point probes were strategically positioned at the system inlet, the exhaust outlet, and along the edge of each tray layer farthest from the air inlet. This conservative placement ensured that the most critical, time-dependent heating curves were accurately monitored.

F. Data Extraction and Evaluation Metrics

To objectively quantify flow equalization and thermal homogeneity, horizontal evaluation planes were established at each tray level. Data extraction was performed across these planes, and the system's performance was evaluated using two dimensionless indices.

The Velocity Uniformity Index (VUI) and Temperature Uniformity Index (TUI) were calculated based on the area-weighted relative standard deviation of the flow fields:

$$\text{VUI} = 1 - \frac{\sigma_v}{\bar{v}} \quad (6)$$

$$TUI = 1 - \frac{\sigma_T}{\bar{T}} \quad (7)$$

where, σ_v and σ_T are the standard deviations of velocity and temperature across the respective evaluation planes, and $\bar{v}$ and $\bar{T}$ represent the corresponding area-weighted mean velocity and temperature. Values approaching 1.0 indicate optimal uniformity across the evaluated plane.

In addition to planar data, discrete monitoring points were utilized for localized analysis. Point probes were strategically placed at the center of the front face of each of the five tray layers. These probes were positioned at the furthest extremity from the air duct inlet to capture the potential "dead zones" or areas of maximum thermal lag, providing a conservative estimate of the drying conditions at the most critical locations within the chamber.

IV. Results

A. Scope and Limitations of the Simulation Study

A full-scale three-dimensional model of the coffee cabinet dryer was analyzed based on its actual geometric dimensions and intended operating parameters. The present investigation is strictly limited to simulation-based performance characterization under no-load conditions.

Since this is a foundational pre-fabrication analysis, key operational metrics heavily dependent on agricultural variables such as drying rate, moisture removal efficiency, specific energy consumption, and final product quality were not experimentally validated. Moreover, this study emphasizes aerothermal performance assessment through validated numerical modeling, establishing the necessary aerodynamic baseline required before future experimental validation and full-load drying kinetics can be quantified.

B. Simulation Performance and Convergence

The segregated solver, employing pseudo-time stepping together with Anderson acceleration, provided stable convergence for the coupled velocity, pressure, turbulence, and heat transfer equations. Convergence for the steady-state domains was achieved within approximately 130 to 170 iterations, with residuals reduced to acceptable tolerances for both fluid flow and turbulence variables. Similar convergence behavior was observed in the plenum and drying chamber simulations. This confirms that the numerical model was capable of predicting steady-state airflow and heat transfer characteristics within the dryer configuration.

The solution time ranged between 20 and 25 minutes for steady-state simulations, while the transient simulations incorporating the piecewise thermostat control logic required approximately 7 to 8 minutes. The convergence profile validates the robustness of the numerical model and supports its suitability for evaluating the aerothermal design of the cabinet prior to physical fabrication.

C. Grid Independence Verification

Across successive mesh refinements, temperature variations remained negligible. The maximum observed deviation between adjacent mesh levels did not exceed 0.09°C, with relative differences remaining below 0.25% of the nominal 40°C (313.15 K) operating temperature. Notably, the heated air inlet temperature remained constant at 40°C across all mesh levels, indicating stable boundary condition implementation.

TABLE II. Grid Independence Results for Predicted Temperature Across All Probes

Mesh Level	Probe A	Probe B	Probe C	Probe D	Probe E	Probe F
h1.0	40.00	39.95	39.93	39.94	39.94	39.94
h1.5	40.00	39.95	39.93	39.95	39.95	39.94
h2.0	40.00	39.95	39.94	39.96	39.95	39.95
h2.5	40.00	39.93	39.90	39.94	39.93	39.92
% Error	**0.00%**	**0.06%**	**0.09%**	**0.05%**	**0.06%**	**0.07%**

As presented in Table II, the temperature values across all six monitoring locations exhibited high stability as the mesh density increased. The primary boundary condition at the inlet (Probe A) remained constant at 40.00°C, confirming that the refinement process did not introduce numerical instabilities at the source. For the internal regions, particularly at the front and center of the third tray (Probes B and F), the temperature variations between successive refinements were minimal, settling at approximately 39.93°C to 39.95°C. The convergence was quantitatively assessed using the relative error between the medium-fine (h2.0) and extremely fine (h2.5) mesh levels. The results indicated that the maximum relative error did not exceed 0.09%, occurring at probe C. In the central drying zones, the error remained below 0.07%.

Based on the numerical stability observed across the mesh iterations, the configuration consisting of 606,877 elements (h2.0) was determined to provide sufficient spatial resolution. The transition to the finest mesh (749,110 elements) yielded a maximum relative error of less than 0.09%, indicating that further refinement offered negligible improvements in precision. Consequently, the h2.0 mesh was selected for all subsequent steady-state and transient simulations to optimize the balance between numerical accuracy and computational efficiency.

Furthermore, no monotonic drift in predicted values was observed with increasing mesh density. Therefore, the mesh containing 606,877 elements was selected as the optimal configuration, effectively balancing computational efficiency and numerical accuracy. This grid independence validation confirms that the velocity, pressure, and temperature fields presented in subsequent sections are strictly mesh-independent and physically reliable.

D. Aerodynamic Performance and Flow Distribution

Achieving a uniform velocity profile across all tray levels is the primary fluid dynamic challenge in multi-tray batch dryers, as non-uniform flow directly induces heterogeneous moisture content [21].

1) Airflow path, velocity uniformity, and intermediate duct dynamics: The inlet boundary condition was set to a velocity of 1.56 m/s at the air duct entry, corresponding to a volumetric flow rate of 0.03 m³/s. The velocity contours (Fig. 5) confirm a peak magnitude of 1.51 m/s within the duct, validating the boundary setup. The high-momentum air builds up in the large plenum chamber, where heating and initial pressurization occur. The conditioned air then exits the plenum and passes through an intermediate air duct before entering the main drying chamber, a configuration intended to redistribute flow pressure and ensure delivery to the base of the tray stack. Within this intermediate duct, the flow velocity is observed to increase due to the reduced cross-sectional area, creating a highly directional jet entering the drying volume.

Within the main drying zone (the volume encompassing the six trays), the air velocity streamlines (Fig. 6) display a highly desirable parallel flow regime as the flow expands and passes through the trays. The velocity magnitude over the virtual coffee bed volume consistently falls within the range of 0.5 m/s to 1.0 m/s. This narrow range indicates a high degree of flow uniformity.

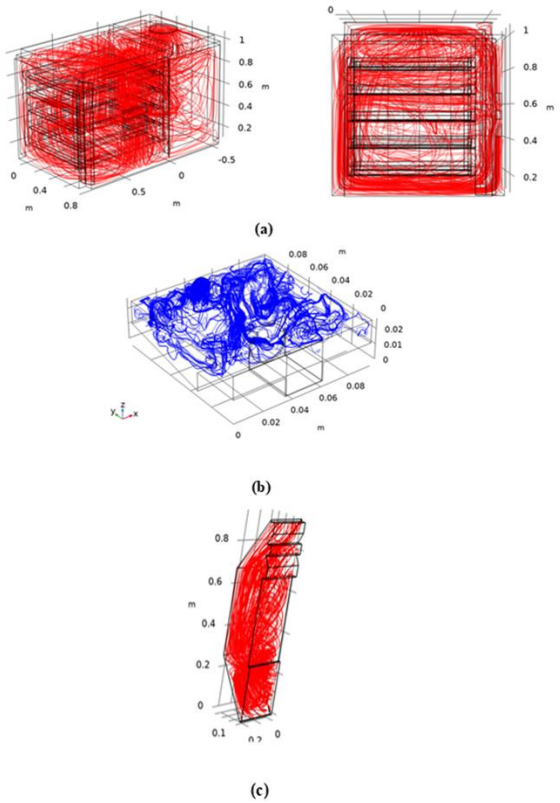


Fig. 5. Streamline visualization of the velocity field in different computational domains: (a) Drying chamber; (b) Air duct; and (c) Plenum chamber.

Assuming an approximate characteristic length (hydraulic diameter) corresponding to the tray spacing, the Reynolds number (Re) calculated for the flow through the drying zone suggests a regime transitioning from laminar to mildly turbulent. This condition is advantageous, as it provides sufficient convective heat transfer without incurring excessive fan power consumption or risk of entraining lighter particles. The minimal localized recirculation zones observed in the far corners of the plenum are low-energy eddies that rapidly decay. The flow entering the intermediate duct is stable, confirming that the initial air distribution within the plenum is effective.

2) Quantitative velocity uniformity assessment: To quantitatively evaluate airflow distribution, the Velocity Uniformity Index (VUI) was computed for each tray surface. The Velocity Uniformity Index (VUI) quantitatively evaluated airflow distribution across the multi-tray system, yielding a global value of 0.84 that indicates highly effective overall balance (Table III).

TABLE III. VELOCITY UNIFORMITY INDEX INSIDE THE DRYING CHAMBER

Location	Individual VUI	Global VUI
Tray No. 1	0.76	0.84
Tray No. 2	0.88	
Tray No. 3	0.87	
Tray No. 4	0.86	
Tray No. 5	0.86	

While Tray 1 showed a modestly lower VUI of 0.76 due to localized jet effects from its proximity to the intermediate duct discharge, upper trays achieved superior uniformity of 0.86 – 0.88 through progressive plenum-driven redistribution, outperforming baseline CFD-optimized designs reporting 0.766–0.80 VUI found in several studies [22][26]. Furthermore, velocities across all evaluation planes remained above stagnation thresholds (>0.5–1.0 m/s), eliminating dead zones and confirming the system's efficacy relative to plenum-enhanced prototypes achieving similar flow homogenization.

Nevertheless, to further improve the aerodynamic uniformity at the lowest tray level prior to physical fabrication, a geometric flow-conditioning modification is proposed at the intermediate duct discharge region. Specifically, integrating a passive flow-correcting element, such as an angled deflector wedge or perforated distribution baffle plate, would mechanically disrupt and diffuse the concentrated high-momentum jet before it impinges on Tray 1. This modification is expected to promote earlier lateral airflow redistribution within the lower plenum region, thereby reducing localized velocity gradients and increasing the Tray 1 VUI toward the more balanced profiles observed in the upper trays.

3) Pressure distribution, hydraulic resistance, and flow equalization: The pressure field (Fig. 7) is perhaps the most compelling result validating the dryer's aerodynamic function. A substantial static pressure peak, exceeding 30 Pa, is successfully established in the plenum immediately adjacent to the air inlet. This pressure acts as the necessary reservoir to

overcome the cumulative hydraulic resistance of the intermediate duct and the subsequent six stacked trays and the coffee product, which will be modeled as a porous medium in subsequent studies.

The principle of flow equalization in multi-tray systems dictates that the pressure drop across the product ($\Delta P_{product}$) must be significantly larger than the velocity pressure variations within the plenum (ΔP_{plenum}) to ensure a consistent flow rate ($\dot{m}$) through each layer. The highly uniform static pressure profile observed beneath the lowest tray suggests that ΔP_{plenum} is negligible compared to the expected $\Delta P_{product}$. The smooth, progressive decrease in pressure moving vertically through the drying chamber, terminating near ambient (zero-gauge) pressure at the exhaust, confirms that the static pressure gradient successfully drives the air across the entire volume. This robust pressure differential fundamentally mitigates the stratification and non-uniform drying characteristic of passively vented systems [20].

E. Thermal Homogeneity

Thermal homogeneity is essential for product quality, particularly for coffee, where temperature excursions above 50°C can initiate irreversible physicochemical damage, compromising final cup quality [16].

1) Thermal mixing and energy balance within the plenum: The simulation employed a high heater outlet temperature of 100°C (373.15 K) as the boundary condition. The temperature gradient reveals a remarkable degree of thermal dissipation within the plenum volume (Fig. 8). The high-velocity, high-temperature air jet interacts immediately with the large, relatively quiescent volume of the plenum. This promotes intense turbulent mixing, causing the air temperature to rapidly reduce and settle into a narrow operational range of 41°C to 45°C (314 K to 318 K) as it enters the intermediate duct.

This rapid stabilization which results in a narrow ±4 K deviation around the target 40°C (313.15 K) confirms effective heat balance. The internal metallic walls and tray structure act as a thermal buffer and heat sink, ensuring thermal energy is distributed uniformly across the plenum's cross-section. The applied external surface heat flux of 180 W/m² successfully counteracts minor conductive losses, preserving the near-adiabatic state of the system. By confirming that the drying air is uniformly applied at a moderate temperature, the simulation provides confidence that the mechanical dryer will preserve the cellular structure, color, and subsequent sensory profile of the coffee.

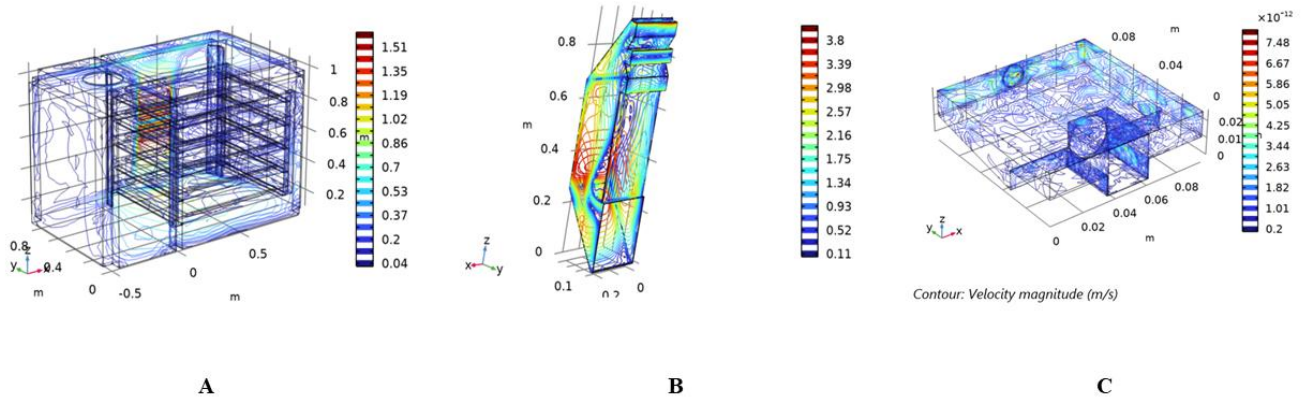


Fig. 6. Contour visualization of the velocity magnitude (m/s): (A) Drying chamber; (B) Air duct; and (C) Plenum chamber.

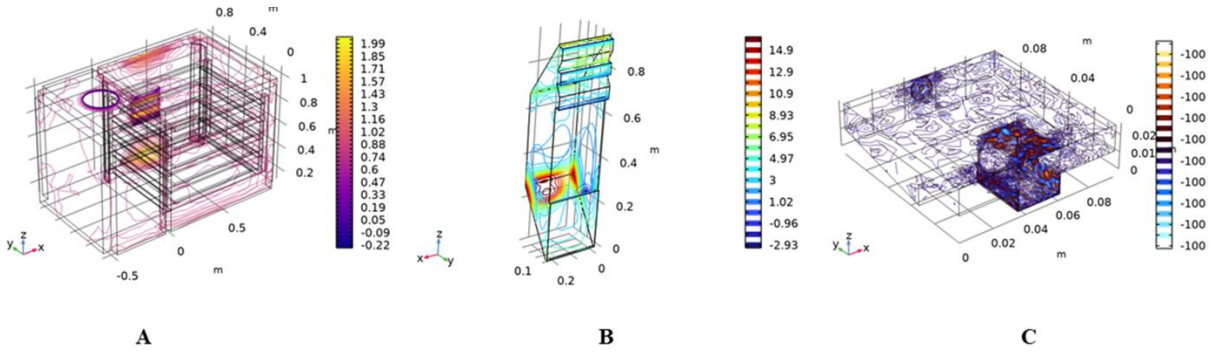


Fig. 7. Contour visualization of the Pressure in Different Domains: (A) Drying chamber; (B) Air duct; and (C) Plenum chamber.

2) Quantitative thermal uniformity assessment: Resulting Temperature Uniformity Index (TUI) demonstrates exceptionally high thermal homogeneity, with a global TUI of 0.99 and individual tray values of 1.00.

While these values approach spatial homogeneity and may appear unusually ideal, they represent the physically coherent thermal behavior of the system under controlled, no-load conditions. The absence of a porous coffee bed eliminates spatially varying permeability and evaporative cooling effects. Under steady-state forced convection with insulated boundaries and intensive plenum mixing, the energy equation naturally converges toward a smooth temperature field without numerical artifacts (Table IV).

TABLE IV. TEMPERATURE UNIFORMITY INDEX INSIDE THE DRYING CHAMBER

Location	Individual TUI	Global TUI
Tray No. 1	1.00	0.99
Tray No. 2	1.00	
Tray No. 3	1.00	
Tray No. 4	1.00	
Tray No. 5	1.00	

This inherent tendency toward thermal equilibration is further reinforced by the aerodynamic configuration of the plenum chamber. Because the heating element is located within a relatively large plenum volume, the incoming high-temperature jet undergoes substantial turbulent mixing before entering the tray stack. The interaction between the hot inlet stream and the plenum walls promotes rapid dissipation of localized thermal gradients and facilitates cross-sectional energy redistribution. Consequently, by the time the airflow reaches the intermediate duct and drying chamber, it has already achieved significant thermal stabilization, thereby limiting the potential for downstream stratification.

The pressure distribution within the system complements this thermal behavior. The establishment of a stable static pressure reservoir in the plenum ensures nearly uniform mass flow across all tray levels. When airflow is evenly distributed and tray geometries are identical, convective heat transfer coefficients remain spatially consistent. This balanced mass flow minimizes localized overheating or underheating, directly supporting the high uniformity reflected in the TUI values. Furthermore, the chamber walls were modeled as thermally insulated, approximating near-adiabatic operation. By limiting asymmetric external heat losses, the model suppresses lateral temperature gradients that could otherwise arise from environmental interactions. The combined effect of insulated boundaries, balanced mass flow, and intensive plenum mixing creates conditions under which temperature deviations are inherently minimal.

Overall, the near-unity TUI values do not indicate numerical artifact or over-idealization; rather, they represent the physically coherent thermal behavior of the system under controlled, no-load conditions. It is expected that the incorporation of porous media modeling and coupled moisture transport in future full-load simulations will introduce additional spatial variability due to evaporative cooling and heterogeneous resistance effects, thereby yielding more conservative yet operationally representative uniformity indices.

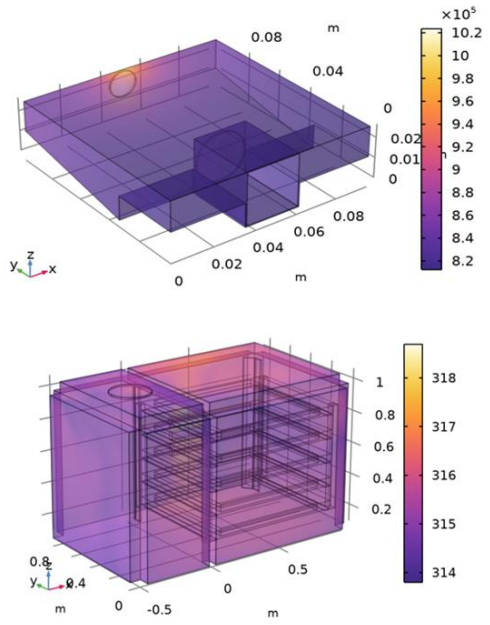


Fig. 8. Temperature distribution within the modeled volume.

F. Transient Thermal Response

In addition to steady-state mapping, a time-dependent simulation was conducted over a 60-minute period. Temperature evolution was monitored using point probes strategically placed at the system inlet, outlet, and along the side of each tray farthest from the air inlet to represent a conservative case for heat delivery. The temporal profiles indicate a rapid rise in air temperature during the initial heating phase, followed by a gradual approach toward steady-state conditions. The inlet and outlet temperature histories show that the system stabilizes within the target operating range after the initial transient period, with minimal temperature divergence observed between the inlet and outlet once equilibrium is reached. This behavior suggests effective thermal regulation and supports the assumption of near-adiabatic operation (Fig. 9).

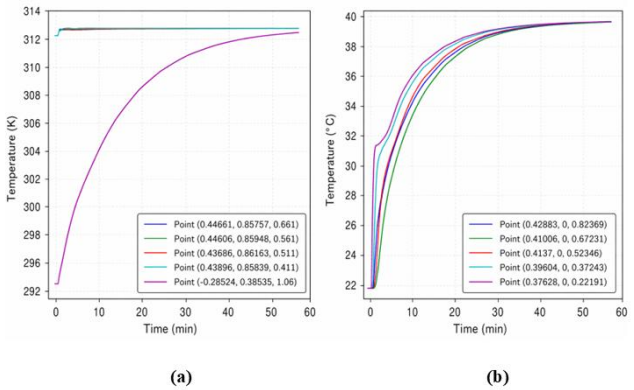


Fig. 9. Transient temperature profiles of the mechanical coffee cabinet dryer obtained from a 60-minute time-dependent CFD simulation. (a) Temperature evolution at selected points near the air inlet and outlet, showing rapid initial heating followed by stability.

Furthermore, the tray-level responses exhibit consistent heating trends across all layers. While minor differences in heating rates are observed during the early stages, all tray locations converge toward a narrow temperature band as time progresses, reaching the desired drying range even at the farthest probe locations.

V. Limitations of the Study

Despite its foundational contributions to the aerothermal design of the cabinet dryer, this study acknowledges several inherent limitations that must be addressed in subsequent research phases. Primarily, the current computational model operates strictly under no-load conditions. The absence of a physical coffee bed means that the complex aerodynamic resistance, spatially varying permeability, and localized flow channeling introduced by agricultural materials are not captured in the current velocity fields. Furthermore, the thermal evaluation is restricted to single-phase sensible heat transfer. In actual operational states, the desorption of moisture from the coffee beans will introduce multiphase mass transfer and significant evaporative cooling effects. These endothermic reactions will inherently induce localized temperature gradients, meaning the near-unity Temperature Uniformity Index (TUI) observed in this baseline study represents an idealized maximum rather than a field-loaded absolute.

VI. Conclusion

The study focused on establishing the aerothermal baseline for the proposed mechanical coffee cabinet dryer through rigorous three-dimensional computational fluid dynamics (CFD) modeling. The results demonstrated that coupling a high-pressure plenum with strategic internal routing has the capability to achieve the flow equalization and thermal stability required for effective batch drying. The aerodynamic evaluation of this architecture revealed a highly stable, parallel flow regime across the drying chamber, evidenced by a global Velocity Uniformity Index (VUI) of 0.84. Airflow velocities were maintained within the target convective range of 0.5 m/s to 1.0 m/s across all product planes, reducing the likelihood of stagnant recirculation zones. Concurrently, the thermal analysis indicated that intensive turbulent mixing within the primary plenum facilitates rapid energy dissipation and stabilization. This mechanism maintained the internal drying environment between 41°C and 45°C (314 K to 318 K), resulting in a high Temperature Uniformity Index (TUI) of 0.99 under the simulated conditions. Such thermal uniformity may help mitigate localized overheating, thereby reducing the risk of irreversible physiochemical damage to the coffee embryo and preserving final cup quality.

However, it must be explicitly emphasized that this near-unity thermal index represents an idealized baseline configuration under empty, no-load conditions. In actual drying operations, the introduction of a physical, porous coffee material will significantly alter these dynamics by introducing spatially varying hydraulic resistance, localized flow channeling, and endothermic mass transfer gradients driven by evaporative cooling. These multi-phase interactions are expected to induce localized temperature gradients, substantially lowering the uniformity indices under full-load scenarios.

Nevertheless, this pre-fabrication characterization provides simulation-based design assessment of the fluid-dynamic and thermal viability of the dryer architecture. By establishing a strictly verified computational framework, these findings secure the aerothermal baseline required for the system's physical implementation. While this study evaluated a specific small-scale dryer architecture, the CFD methodology and geometric principles of flow equalization demonstrated herein may be adapted to optimize various types of agricultural drying equipment.

Acknowledgment

This work was supported by the Cavite State University (CvSU) and the Department of Science and Technology (DOST) through the DOST-funded project titled "Technology Piloting and Commercialization of Coffee Postharvest Equipment" under the program titled "Bridging the Gap on Coffee Production and Processing through Emerging Research and Innovation for the Advancement of the Industry (CvSU BEAN Program)." The authors also express their sincere gratitude to the coffee farmers, producer organizations, Local Government Units, and partner communities across CALABARZON for their invaluable support and cooperation throughout the conduct of this work.

References

[1] A. K. Stephen and S. Emmanuel, "Improvement on the design of a cabinet grain dryer," *American Journal of Engineering and Applied Sciences*, vol. 2, no. 1, pp. 217–228, 2009, doi: 10.3844/AJEAS.2009.217.228.

[2] İ. Doymaz and M. Pala, "The thin-layer drying characteristics of corn," *Journal of Food Engineering*, vol. 60, no. 2, pp. 125–130, 2003, doi: 10.1016/S0260-8774(03)00016-0.

[3] O. Durodola, "Design and performance evaluation of a multipurpose conductive rotary dryer adapted for drying ogi," 2022. doi: 10.13140/RG.2.2.27692.33923.

[4] P. Corrêa, O. Resende, and D. Ribeiro, "Drying characteristics and kinetics of coffee berry," *Revista Brasileira de Produtos Agroindustriais*, vol. 8, no. 1, pp. 1–10, 2006.

[5] E. Menya and A. Komakech, "Investigating the effect of different loading densities on selected properties of dried coffee using a GHE dryer," *Agricultural Engineering International: CIGR Journal*, vol. 15, no. 2, pp. 231–237, 2013.

[6] P. Ghosh and N. Venkatachalapathy, "Processing and drying of coffee – a review," *International Journal of Engineering Research and Technology*, vol. 3, no. 12, pp. 784–794, 2014.

[7] B. Soeswanto, N. Wahyuni, and G. Prihandini, "The development of coffee bean drying process technology – a review," in *Proceedings of the International Conference on Agricultural Engineering and Rural Development*, 2021, doi: 10.2991/aer.k.211106.026.

[8] A. W. Aregba, P. Sebastian, and J. Nadeau, "Stationary deep-bed drying: A comparative study between a logarithmic model and a non-equilibrium model," *Journal of Food Engineering*, vol. 77, no. 1, pp. 27–40, 2006, doi: 10.1016/j.jfoodeng.2005.06.020.

[9] C. L. Oria and E. V. Palconit, "A review on possible combination of solar dryer materials for crops in the Philippines," *Engineering and Technology Journal*, vol. 7, no. 7, pp. 1328–1341, 2022.

[10] E. Duque-Dussán, J. R. Sanz-Uribe, and J. Banout, "Design and evaluation of a hybrid solar dryer for postharvesting processing of parchment coffee," *Renewable Energy*, vol. 215, p. 118961, 2023, doi: 10.1016/j.renene.2023.118961.

[11] E. M. Meja, S. K. Dubbe, A. Bekele, K. F. Wolde, and M. S. Adaramola, "Investigating the performance and optimization of solar coffee drying technologies: A systematic review," *Journal of Food Processing and Preservation*, vol. 49, no. 1, 2025, doi: 10.1111/jfpp.70028.

[12] R. G. Idago and R. S. M. Dela Cruz, *Value Chain Improvement of Robusta and Liberica Coffee*. Science City of Muñoz, Nueva Ecija, Philippines: Philippine Center for Postharvest Development and Mechanization (PhilMech), Tech. Bull., vol. 5, no. 1, 2015.

[13] A. Parra-Coronado, G. Roa-Mejía, C. E. Oliveros-Tascón, and J. R. Sanz-Uribe, *Optimización operacional de secadores mecánicos para café pergamino*. Chinchiná, Colombia: CENICAFÉ, 2017.

[14] S. Phitakwinai, S. Thepa, and W. Nilnont, "Thin-layer drying of parchment Arabica coffee by controlling temperature and relative humidity," *Food Science and Nutrition*, vol. 7, no. 9, pp. 2921–2931, 2019, doi: 10.1002/fsn3.1142.

[15] R. A. Trejos, G. Roa-Mejía, and C. E. Oliveros-Tascón, "Humedad de equilibrio y calor latente de vaporización del café pergamino y del café verde," *Revista Cenicafé*, vol. 40, no. 1, pp. 5–15, 1989.

[16] R. Osorio Hernandez, I. D. F. Ferreira Tinoco, J. Correna Carlo, J. A. Osorio Saraz, and I. D. Aristizábal Torres, "Bioclimatic analysis of three buildings for wet processing of coffee in Colombia," *Revista Facultad Nacional de Agronomía Medellín*, vol. 71, no. 3, pp. 8609–8616, 2018, doi: 10.15446/rfnam.v71n3.67772.

[17] M. B. Fauzi, E. A. Kosasih, M. I. Dzaky, G. G. Nasution, and A. Zikri, "Experimental analysis and numerical simulation of coffee bean drying in packed beds with different coffee bean stack height," *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences*, vol. 128, no. 2, pp. 122–138, 2025, doi: 10.37934/arfmts.128.2.122138.

[18] D. J. M. de Abreu et al., "Influence of drying methods on the post-harvest quality of coffee: Effects on physicochemical, sensory, and microbiological composition," *Foods*, vol. 14, no. 9, p. 1463, 2025, doi: 10.3390/foods14091463.

[19] S. Misha, S. Mat, M. H. Ruslan, K. Sopian, and E. Salleh, "The prediction of drying uniformity in tray dryer system using CFD simulation," *International Journal of Machine Learning and Computing*, vol. 3, no. 5, pp. 419–423, 2013, doi: 10.7763/IJMLC.2013.V3.352.

[20] C. J. Ohagwu et al., "Computational fluid dynamic model analysis of multipurpose dryer with bio-waste heat source: An experimental validation using paddy rice," *Energy Conversion and Management X*, vol. 23, p. 100652, 2024, doi: 10.1016/j.ecmx.2024.100652.

[21] D. Cebulski and P. Cyklis, "Application of CFD simulation to the design of an innovative drying chamber," *Energies*, vol. 17, no. 13, p. 3338, 2024, doi: 10.3390/en17133338.

[22] D. A. Asrate, F. B. Tadle, and Y. D. Mihretie, "Enhancement of airflow uniformity in multi-tray dryers through a CFD-integrated sequential linear programming framework," *Scientific Reports*, vol. 16, p. 4226, 2026, doi: 10.1038/s41598-025-34360-1.

[23] L. Kumar et al., "Experimental and computational validation of thermal performance of an active greenhouse solar dryer in no-load conditions," *International Journal of Low-Carbon Technologies*, vol. 19, p. ctae036, 2024, doi: 10.1093/ijlct/ctae036.

[24] A. Jha and P. P. Tripathy, "Performance evaluation and finite element modeling of heat, mass, and fluid flow inside a hybrid solar dryer during drying of paddy grains," *Frontiers in Food Science and Technology*, vol. 4, p. 1411956, 2024, doi: 10.3389/frfst.2024.1411956.

[25] S. Singh and S. Kumar, "Testing method for thermal performance-based rating of various solar dryer designs," *Solar Energy*, vol. 86, no. 1, pp. 87–98, 2012, doi: 10.1016/j.solener.2011.09.009.

[26] M. Precoppe, S. Janjai, B. Mahayothee, et al., "Batch uniformity and energy efficiency improvements on a cabinet dryer suitable for smallholder farmers," *Journal of Food Science and Technology*, vol. 52, no. 8, pp. 4819–4829, 2015, doi: 10.1007/s13197-014-1544-y.

Unveiling Temporal Dynamics of Consumer Emotions in Online Reviews: An Emotion Mining Approach Using Ekman's Emotion Model

Azra Shamim, Abdulwahab Ali Almazroi
College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia

***Abstract*—Currently, business organizations are using Electronic Word-of-Mouth (EWOM) from consumer-opinion platforms to elevate their marketing strategies using semantic analysis and emotion mining. Existing efforts are devoted to analyzing EWOM using fine-grained emotion classification; however, the well-known and well-established emotion models that provide a basic set of psychological and biological emotions are not targeted in these studies. Further, the existing literature is unable to present temporal dynamism and volatility of consumers' emotions according to emotion categories that have an inescapable impact on consumers and corporate decision-making process. Therefore, this study tries to address this gap by extracting, classifying, summarizing, and tracking consumers' emotions for corporate and consumers' decision-making processes based on widely used and well-established Ekman's emotions model. Therefore, the current work aims to provide consumers' emotions temporal dynamism and volatility at product and feature levels by using binary and fine-grained emotion classifications based on Ekman's basic psychological and biological emotions. Online reviews from amazon.com are used for experimental purposes. The results of the study exhibited that consumers predominantly expressed joy emotion to the camera. However, the picture feature of the camera evoked negative emotions, namely, fear, anger, and sadness. Temporal analysis unveiled shifting emotional expressions in quarter 1 and quarter 2. The results of the study will assist corporates in shaping their decision-making process and marketing strategies. Moreover, the findings of the study provide valuable insights into consumers to improve their purchase decisions by pinpointing the advantages and disadvantages of critical product features through emotional clues.**

Keywords*—Emotion; emotion mining; semantic analysis; online reviews

I. Introduction

The proliferation of online review platforms has transmuted content publishing from corporates towards consumers [1]. The web offers highly scalable and accessible platforms to consumers for the creation and dissemination of Electronic Word-of-Mouth (EWOM) such as wikis, discussion forums, tweets, blogs, and online review websites. EWOM is a positive or negative statement about an object such as product, people, service, topic, organization, set by the potential, actual or former consumers which is available online to a vast audience [2]. Currently, consumers are connected globally to online platforms to express their emotions, evaluations, opinions, and judgments to a broad spectrum of individuals [3][4]. Therefore, fast-growing EWOM provides unstructured, trustworthy, and freely available valuable decision-oriented data that can address the diverse needs of corporates and consumers by providing insights into consumers' behaviour, and preferences[5][6][7].This powerful source of emotions leverages corporates in various business intelligence tasks, namely, sale prediction, risk analysis, enterprise strategies, and benchmarking. On the other hand, it assists consumers in their purchase decisions by highlighting the advantages and disadvantages of a target product [8][9].

Although there are many online platforms for consumers to express their opinions, evaluations, and emotions; however, online reviews are considered more credible and trustworthy [10]. Further, the characteristics of Web 2.0, such as scalability, accessibility, and enhanced users' participation, allow consumers all over the world to express their emotions by writing reviews that result in the explosive growth of online reviews [11]. The massive freely available online reviews present valuable insight into critical features of a target product that play an important role in consumers' purchases and business strategies [7][12][13]. Therefore, corporates are targeting consumers' online reviews from different sources for their decision-making process [14][15]. Major retailers' websites provide a star rating of a product and the helpfulness of the review such as amazon.com, ebay.com, and aliexpress.com. On the other hand, cnet.com presents the rating of a product and its prominent features. However, the emotions expressed by consumers are not considered and published by these websites. Further, it is difficult to extract consumers' emotions manually on critical features of a product due to the volume, distribution, length, structure, multidimensional data and dynamic nature of the online reviews[7][16].

Early efforts and attempts of opinion mining focused on the classification of online reviews into positive or negative opinion categories [17][18]. Later, online reviews are classified into fine-grained opinion categories [19][20][7]. However, the fine-grained opinion categories are unable to provide a perception of consumers' emotions. The classification of EWOM into emotion categories may provide an enhanced understanding of consumers' state of mind, evaluation, and judgment about a particular product. Further, limited attempts are also performed to provide fine-grained emotion classification of online reviews that provide a deeper understanding of consumers' opinions, sentiments, emotions, and evaluations about a target product or feature. However, the existing studies do not classify emotions into well-known and well-established basic psychological and biological emotions as there is a limited number of emotions that

are biologically and psychologically basic to human beings [21]. Further, consumers' emotions may change with time, that is, positive emotions change into negative emotions and vice versa as exhibited by Twitter data [22]. Existing studies [23][24]. are not reporting this temporal dynamism and volatility of consumers' emotions according to emotion categories that have an enormous impact on corporate and consumers' decision-making process. Therefore, it is vital to track this shift in the emotions of consumers of a particular feature or a product over time based on psychological and biological emotions, and the reason behind this change should be identified to make appropriate decisions. Consequently, there is a need to extract, classify, summarize, and track consumers' emotions according to basic psychological and biological emotions for corporate and consumer decision-making.

The aim of this work is to extract, classify, and summarize consumers' emotions expressed in online reviews, and to present the temporal dynamism and volatility of consumers' emotions according to well-established and widely used basic psychological and biological emotions [25] that impacts corporate planning and the consumers' decision-making process significantly. Specifically, the current study 1) classifies and summaries consumers' online reviews into binary and fine-grain classes according to the well-known psychological and biological emotions at the product level, 2) classifies and summaries consumers' online reviews into binary and fine-grain classes according to the well-known psychological and biological emotions at the feature level, 3) provides temporal dynamism and volatility of consumers' emotions according to psychological and biological emotions at the product level, and 4) provides temporal dynamism and volatility of consumers' emotions according to psychological and biological emotions at the feature level. A dataset consisting of real online reviews from amazon.com is used for the experimental purpose in this study.

This study has the following contributions:

- The classification of consumers' emotions into well-known six basic psychological and biological emotions categories: anger, fear, disgust, joy, surprise, and sadness defined by Ekman [25] is presented by the study.
- The study highlights six basic psychological and biological consumers' emotions for target products that show the strengths and weaknesses of target products reported by the actual users of products.
- The six basic psychological and biological consumers' emotions for prominent features of target products are also reported by the study that can play a decisive factor in the purchase decision of consumers.
- The temporal dynamism and volatility of six basic psychological and biological consumers' emotions are observed and presented by the study that is helpful to understand the change in consumers' emotions over time. The temporal dynamism of consumers' emotions is highlighted at two levels: product and feature in the study.

The rest of the paper is organized as follows: Existing emotion models and related work are reviewed in Section II. Section III discusses the proposed methodology. Section IV described the experimental results and discussion. Section V presents the research implications. Lastly, Section VI concludes the paper.

II. Literature Review

A. Emotion Models

Emotions are strong positive or negative motivational stimuli that comprise experiences such as love, hate, anger, joy, and grief. To study the emotions expressed by humans, various emotion models are presented in the literature of Psychology. And these emotion models are a critical component of emotion mining systems [26]. Therefore, this section reviews the state-of-the-art emotion models by focusing on key emotion categories and their associated descriptors. These descriptors play a crucial role in the measurement of emotions. Moreover, basic emotions are the motions having no other emotional constituent parts. Table I provides a summary of existing emotion models.

TABLE I. Existing Emotion Models

Reference	Basic Emotions
[27]	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
[28]	Fear, love, rage
[29]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
[30]	Pain, pleasure
[31]	Joy, Anguish, Fear, Anger, Disgust, Surprise, Interest, Shame
[25]	Joy, Sadness, Fear, Anger, Disgust, Surprise
[32]	Interest, joy, surprise, sadness, anger, disgust, contempt, self-hostility, fear, shame, shyness, guilt
[33]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
[34]	Expectancy, fear, rage, panic
[35]	Happiness, sadness
[36]	Desire, happiness, interest, surprise, wonder, sorrow
[37]	Love, Joy, Surprise, Anger, Sadness, Fear
[31]	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
[38]	Fear, Worry, Anger, Joy, Sadness, Discontent, loneliness, Peacefulness, Excitement, Optimism, Surprise, Shame, Guilt, Romantic Love, Love, Contentment, Envy
[39]	Fear, Anger, Enjoyment, Distress, Disgust, Surprise, Contempt
[40]	contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame

While researchers and psychologists have proposed various emotion models, however, Ekman's model [25] with six basic emotions is considered the most widely accepted and used model in computer science and natural language processing research areas [26][41][42][43]. This model presents six basic emotions: anger, fear, disgust, joy, surprise, and sadness. Later, Ekman expanded these basic emotions by introducing additional positive and negative emotion dimensions [40]. Due to widespread adoption and validation of Ekman's basic emotion model, the current study utilized this model to classify the

emotion expressed by consumers into the six basic emotion categories.

B. Related Work

Emotion mining has been explored in various contexts, including blogs [41], health [42], social media platforms like Facebook [43] and Twitter [44], news headlines [45] [46], online reviews [21], customer emails [47], and stock exchange [48]. These applications aim to extract and analyze emotions expressed by individuals or groups, providing valuable insights into human behavior and sentiment.

Significant work in the area of emotion classification is performed on Twitter data to analyze emotional content in tweets. Recently, [22] classified the public tweets into binary classes as well as into the basic emotions. Similarly, [49] exploited machine learning techniques for the classification of tweets in basic emotion categories. Further, the work analyzed the changes in emotions over time. Likewise, the artificial intelligence methods are utilized in [50] to analyze 73,000 public conversations on Twitter associated with the Corona pandemic using basic emotions proposed by [51]. [52] classified 9,000,000 tweets from a specific time frame into six moods: tension, anger, depression, vigor, confusion, and fatigue using the proposed profile of mood states method. On the other hand, [53] utilized hashtags of tweets as emotion labels for the classification of emotions expressed in tweets. A linear classifier is used to classify 2,500,000 tweets using hashtags as emotion labels into seven emotions: anger, joy, fear, love, sadness, thankfulness, and surprise. The comparison of manually labeled tweets and hashtags labels showed acceptable agreement. Further, the result of experiments conducted using different features like unigrams, bi-grams, POS tags, adjectives, and lexicons exhibited that unigrams, bi-grams, and POS tags provide the best emotion detection results. Moreover, the findings of the study suggested that the increased size of the training dataset has a direct impact on the accuracy of the classifier. In [54], the authors also used hashtags to classify a dataset of 134,000 tweets. They compared hashtags with manual tags assigned by two groups: a group of people and a group of psychologists. The findings of [53] [54] suggest that larger training datasets and features like uni-grams, bi-grams, and POS tags can improve emotion detection accuracy.

In the context of blogs, [55] [56] classified online diary entries into 1 of 40 moods of the writer using a support vector machine. Mood classification is considered similar to emotion classification; however, emotions are more intense in nature. The previous work was enhanced by [38] by capturing and aggregating the moods of writers from a large number of blog posts in a certain period. Similarly, [46] analyzed the emotional content in online news.

In [57], the authors identified emotion passages from the children's fairy tales dataset and then classified these emotions into positive and negative valence. The results from their linear classifier (Sparse Network of Winnows) are not very promising; however, their dataset is extensively utilized by the emotion mining community. Fine-grained classification of emotions from text is performed by [58]. In this work, the authors targeted the 600 notes of persons who committed suicide and predicted a label for each note from 15 emotions such as love, hopelessness, thankfulness, sorrow, and anger. Another work on fine-grained emotion classification of text is presented in [59]. The authors classified the emotions from a pre-defined emotion set that includes six emotions: Joy, Fear, Anger, Disgust, Sadness, and Surprise using a Vector Space Model [60].

In the context of consumers' emotions in online reviews, [61] explored the relationship between three emotion classes (positive, negative, and mixed emotions) with product attitude. On the other hand, [62] identified Plutchick's basic emotions [51] by using the NRC emotion lexicon in online movie reviews from IMDb. Unlike previous studies, [62] analyzed negative emotions in online reviews in a different perspective by exploring the association between two negative emotions (regret and frustration) with the helpfulness of review, reviewer rationality, and price fairness in online review and their empirical results confirmed a positive relationship between these negative emotions and helpfulness of review and price fairness, however, the positive impact of the negative emotions on reviewer rationality was not established. The authors in [47] focused on the customer care domain by identifying customers' emotional emails complaining about services or products to provide timely and pertinent responses to the sender of emotional emails in order to increase customer satisfaction and retention. They used the boosting method and performed experiments on 1077 customers' emails. Their method, based on salient features, outperforms the baseline method based on the unigram feature. [63] analyzed the correlation between helpfulness, unhelpfulness, rating, positive score, and the negative score of product reviews from amazon.com. The impact of the emotional score is not found to be significant on helpfulness in this study; however, a negative correlation is established between emotional score and rating. The emotional score is divided into positive and negative; positive emotion score (1 to 5), whereas negative emotion score (-1 to -5). The distribution of positive and negative scores over time is also highlighted in the study. In contrast, [64] found an association between emotional score and helpfulness of a review. The existing emotion models are not applied in this work for the extraction of emotions. Further, [65] classified emotions expressed in online reviews into positive, negative, and neutral categories. Also, [66] classified online reviews into positive, negative and neutral sentiment using machine learning techniques. Moreover, binary emotion classification is considered by [67][68]. On the other hand, in [69] [70], the authors focused on positive, negative, and neutral emotions. Different experiments were conducted to investigate the link between emotional expressions in reviews and their impact on product evaluation in [71]. [72] investigated the relationship between additional online reviews and sales volume. [21] presented the distribution of positive and negative emotions over time.

Extensive literature has been available targeting 'emotions' in various contexts, like Twitter, online news, and blogs. According to the best of our knowledge, little work has been conducted to extract fine-grained emotion classification according to well-known and widely used six basic psychological and biological emotions in the context of amazon.com at product and feature levels. Moreover, the temporal dynamism and volatility of consumers' emotions are

not tracked by existing work according to basic psychological and basic emotions. Therefore, the current work focuses on the extraction, classification, and tracking of basic psychological and biological emotions: anger, fear, disgust, joy, surprise, and sadness, proposed by [22] from online reviews of amazon.com.

III. Research Design

A. Research Framework

Let document D contain n reviews $R = [r_1, r_2, r_3, \ldots\ldots, r_n]$. Each review r_k is represented by a tuple having two elements: metadata (MD_{r_k}) and body (B_{r_k}) of the review (r_k). This study extended the tuple of [23]. This tuple is termed as the review tuple. The tuple is shown in Fig. 1 and is presented as follows:

$$r_k = [MD_{r_k}, B_{r_k}]$$

The metadata of the review (r_k) is decoded as MD_{r_k} and is comprised of a review rating ($Rating_{MD_{r_k}}$), review title ($Title_{MD_{r_k}}$) and date ($Date_{MD_{r_k}}$) whereas B_{r_k} signifies a set of sentences in r_k as shown below:

$$MD_{r_k} = [Rating_{MD_{r_k}}, Title_{MD_{r_k}}, Date_{MD_{r_k}}]$$

$$B_{r_k} = [S_1, S_2, S_3, \ldots, S_m]$$

$$\forall\, S_i \in B_{r_k}, S_i = [B_{r_k} S_i F_j, B_{r_k} S_i F_j ECat, B_{r_k} S_i F_j E_w, B_{r_k} S_i F_j E, B_{r_k} S_i Content]$$

Every sentence S_i in B_{r_k} encloses a feature ($B_{r_k} S_i F_j$), emotion category of emotion ($B_{r_k} S_i F_j ECat$), emotion word ($B_{r_k} S_i F_j E_w$), emotion expressed on F_j ($B_{r_k} S_i F_j E$) and contents of S_i ($B_{r_k} S_i Content$).

Every review (r_k) embodies a set of feature-emotion pairs (F_j, E_w). This pair includes a feature (F_j) and associated emotion word (E_w). Every feature F_j may have more than one associated E_w in document *D*.

The emotion expressed in the feature F_j in sentence S_i classified into two categories: positive ($B_{r_k} S_i F_j ECat^{Pos}$) or negative ($B_{r_k} S_i F_j ECat^{Neg}$). In other words, $B_{r_k} S_i F_j ECat$ is either positive or negative. $B_{r_k} S_i F_j ECat^{Pos}$ have two positive emotions: Joy ($B_{r_k} S_i F_j ECat^{Pos_Joy}$) and surprise ($B_{r_k} S_i F_j ECat^{Pos_Surpise}$). On the other hand, anger ($B_{r_k} S_i F_j ECat^{Neg_anger}$), fear ($B_{r_k} S_i F_j ECat^{Neg_Fear}$), disgust ($B_{r_k} S_i F_j ECat^{Neg_Disgust}$) and sad ($B_{r_k} S_i F_j ECat^{Neg_Sad}$) are negative emotions that belong to $B_{r_k} S_i F_j ECat^{Neg}$.

$$B_{r_k} S_i F_j ECat = [B_{r_k} S_i F_j ECat^{Neg}, B_{r_k} S_i F_j ECat^{Pos}]$$

$$B_{r_k} S_i F_j ECat^{Pos} = [B_{r_k} S_i F_j ECat^{Pos_Joy}, B_{r_k} S_i F_j ECat^{Pos_Surpise}]$$

$$B_{r_k} S_i F_j ECat^{Neg} = [B_{r_k} S_i F_j ECat^{Neg_anger}, B_{r_k} S_i F_j ECat^{Neg_Fear}, B_{r_k} S_i F_j ECat^{Neg_Disgust}, B_{r_k} S_i F_j ECat^{Neg_Sad}]$$

Table II illustrates the description of notations used in the proposed review tuple.

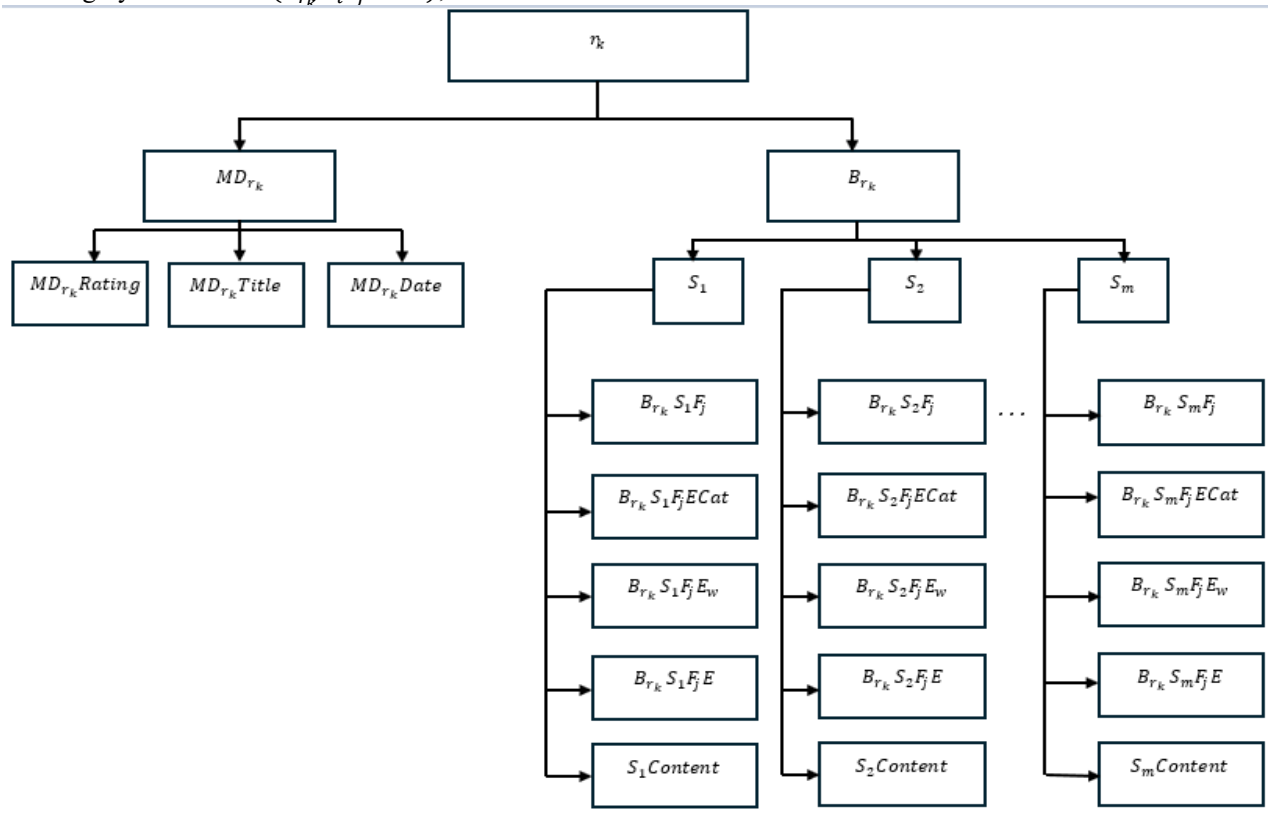


Fig. 1. Proposed tuple.

TABLE II. DESCRIPTION OF NOTATIONS

Notations	Description
D	Document having n product reviews
r_k	Review k
MD_{r_k}	Metadata of Review k
B_{r_k}	Body of Review k
$MD_{r_k}Rating$	Rating of Review k in the metadata of review
$MD_{r_k}Title$	Title of Review k in the metadata of review
$MD_{r_k}Date$	Date of Review k in the metadata of review
$B_{r_k}\ S_iF_j$	A product feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat$	Emotion category (Positive/Negative) of emotion word E_w associated with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jE_w$	Emotion word E_w associated with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jE$	Emotion E associated with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iContent$	The content of S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Pos}$	Indicates emotion category of emotion word E_w associated with feature (F_j) in sentence S_i in body of review k is Positive
$B_{r_k}\ S_iF_jECat^{Neg}$	Indicates emotion category of emotion word E_w associated with feature (F_j) in sentence S_i in body of review k is Negative
$B_{r_k}\ S_iF_jECat^{Pos_Joy}$	Indicates Joy emotion expressed on with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Pos_Surpise}$	Indicates Surprise emotion expressed on with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Neg_anger}$	Indicates Anger emotion expressed on with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Neg_Fear}$	Indicates Fear emotion expressed on with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Neg_Disgust}$	Indicates Disgust emotion expressed on with feature (F_j) in sentence S_i in body of review k
$B_{r_k}\ S_iF_jECat^{Neg_Sad}$	Indicates Sad emotion expressed on with feature (F_j) in sentence S_i in body of review k

B. System Architecture

The architecture of the proposed system is discussed in this section. A system called Emotion Analyzer is developed in this work to provide fine-grained emotion classification at product and feature levels, and to track temporal dynamism and volatility of consumers' emotions. The system is comprised of four components: data preprocessor, feature and emotion extractor, emotion classifier, and emotion visualizer, as shown in Fig. 2 and discussed below.

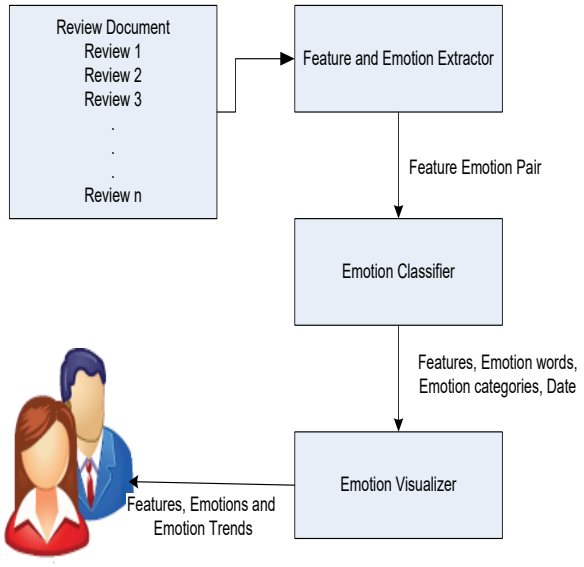


Fig. 2. Proposed system.

1) Data pre-processor: Data pre-processor performs a variety of operations to formulate document D containing n reviews for feature and emotion extraction. The following operations are performed by the component. Data processing is performed to increase classification accuracy.

- The symbols, special characters, and punctuations are removed from document *D*.
- Stop words are discarded from document *D*.
- The inflected or derived words are converted to their root form by utilizing Word stemming to enhance the efficiency and accuracy of mining.
- Spell checking is performed to eliminate noise from document *D*.
- Part-of-speech tagging is applied to label each word with its corresponding category, i.e., noun, verb, and adjective.

2) Feature and emotion extractor: Feature and emotion extractor identifies critical product features with their associated emotion words. The input to feature and emotion extractor is document D consisting of n reviews of a target product. A file containing all feature-emotion pairs is the output of feature and emotion extractor. In order to create this file, in the first step, the decisive features of a target product are extracted by the feature and emotion extractor. The candidate's features are derived from frequent nouns. All frequent nouns are obtained from document D and stored in a list called

Nounlist[]. Then, a weight is assigned to each noun in Nounlist [].

Existing literature presents a multitude of schemes for feature weight calculation, such as word frequency-inverse document frequency (TF-IDF), chi- square test, and information gain. However, among the aforementioned schemes, TF-IDF is regarded as the most ubiquitous and efficient scheme since it can reflect the importance of words within a collection of documents [79]. Therefore, the current work utilized and enhanced this scheme for feature weight calculation. The formula of TF-IDF utilized in the current study is shown in Eq. (1) – (3).

$$f_j f(f_j, r_k) = \frac{f_{f_j, r_k}}{\sum f_j' \in r_k f f_j', r_k} \quad (1)$$

$$idf(f_j, D) = Log\ (\frac{|R|}{n_{f_j}}) \quad (2)$$

$$f_j f.idf(f_j, r_k, D) = \frac{f_{f_j, r_k}}{\sum f_j' \in r_k f f_j', r_k} * Log\ (\frac{|R|}{n_{f_j}}) \quad (3)$$

In Equation 1, f_{f_j, r_k} is the number of times f_j occurs in r_k and is calculated using Eq. (4).

$$f_{f_j, r_k} = \sum_{i=1}^{n} Count\ f_j\ occurence\ in\ r_k \quad (4)$$

The denominator in Equation 1 shows the total number of features in r_k.

$|R|$ in Eq. (2) shows the total number of reviews in D. $n_{f_j} = |\{r_k \in D : f \in r_k\}|$: number of reviews in which the f_j appears where $\text{tf}(t, d) \neq 0$ and computed using equation below:

$$n_{f_j} = \sum_{i=1}^{n} [f_j Existin r_k = True] \quad (5)$$

Eq. (6) and (7) compute the weight of features by introducing a new variable into TF-IDF calculation, that is, the number of emotion words associated with f_j $(f_j E_w Count)$. The introduction of the new variable rests on the assumption that more emotions expressed toward a feature indicate frequent discussion of that feature, thus justifying a higher relative weight compared to other features.

Eq. (8) exhibits the computation of $f_j E_w Count$.

$$f_j W = f_j f.idf(f_j, r_k, D) + f_j E_w Count \quad (6)$$

$$f_j W = \frac{f_{F_j, r_k}}{\sum f_j' \in r_k f f_j', r_k} * Log\left(\frac{|R|}{n_{f_j}}\right) + f_j E_w Count \quad (7)$$

$$f_j E_w Count = \sum_{k=1}^{n} Count\ E_w assoiated\ with\ f_j(r_k) \quad (8)$$

Eq. (7) is used to compute the weight of a feature f_j. Nouns in NounList[] having $f_j W$ above a threshold α are considered as potential features as shown in Eq. (9).

$$F = [f_1, f_2, f_3, \ldots\ldots, f_n] \quad (9)$$

After this, E_w associated with each feature are extracted by exploiting a window-based approach in which E_w conversed within K words of f_j are selected as associated emotion words ($f_j E_w$). At the end, a feature emotion list is created that contains all features and associated emotion words as shown below:

If f_j is a feature in the F, then $f_j E_w = \{F_j E_{W_1}, F_j E_{W_2}, \ldots F_j E_{W_n}\}$. The feature-opinion pairs (FW) are stored in a ragged array represented below:

FW ∈ ℝ^{n × m_i} where $n = \sum_{-}\{x \in F\}$ 1 and m_i is the number of column in i[th] row.

3) Emotion classifier: The task of emotion classification is performed by the emotion classifier. The input to the emotion classifier is FW and the output is the classification of emotions. Extensive emotion mining systems rely on emotion lexicon (dictionary-based approach) for the classification of emotions, and some emotion lexicons are available in the literature for the classification of emotions, such as Wordnet Affect [74]. The Linguistic Inquiry and Word Count (LIWC) [75], NRC emotion lexicon [76], Clean Balanced Emotional Tweets (CBET)[76] [77]. Recently, [78] utilizes the NRC lexicon to classify public tweets into the emotion categories of Ekman. Therefore, the current study also utilized the NRC emotion lexicon [73].

Given the feature-emotion pairs FW, the emotion classification function f classifies $F_i W_j$ to an emotion category e_k using lexicon L.

Formally, f can be defined as:

$$f \triangleq (FW_{ij}, L) \mapsto L(e_{ij}) = e_k$$

where,

$$FW_{ij} = \left(f_i, E_{W_j}\right) \in FW, E_{W_j} \in E_w, L\left(E_{W_j}\right) = e_k \in E$$

$$FW_{ij} \subseteq F \times E_w$$

L ∈ ℝ^(n×m), where n is the number of words in the lexicon and m is the number of emotion categories.

$E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ where e_1 =Joy, e_2 =surprise, e_3=anger, e_4=fear, e_5=disgust, e_6=sad.

In other words,

f: $FW_{ij} \rightarrow$ E, where $g(E_w) = \text{argmax_e} \in E\ (P(e|E_w))$, and $P(e|E_w)$ is the probability of emotion e given the word E_w

The classified emotions are aggregated at the feature using the following equations:

$$Countf_j e_1 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i\ f_j E_w = e_1]$$

$$Countf_j e_2 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i\ f_j E_w = e_2]$$

$$Countf_j e_3 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i\ f_j E_w = e_3]$$

$$Countf_j e_4 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i\ f_j E_w = e_4]$$

$$Countf_j e_5 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i f_j E_w = e_5]$$

$$Countf_j e_6 = \sum_{r=1}^{m} \sum_{i=1}^{sentence(r)} [S_i f_j E_w = e_6]$$

The classified emotions are then aggregated at the feature and product levels using the following equations:

Feature-level aggregation: F_e = ∑_{w∈W} P(e|w) / |W|

Product-level aggregation: P_e = ∑_{f∈F} F_e / |F|

where, F_e is the feature-level emotion score, P_e is the product-level emotion score, and |W| and |F| are the number of words and features, respectively.

The emotion classifier provides a file to the emotion visualizer that consists of a list of features, associated emotion words, emotion categories, and the date on which this emotion category was expressed by the consumer. This can be represented as a tuple (F, W, E, D), where F is the set of features, W is the set of emotion words, E is the set of emotion categories, and D is the set of dates.

4) Emotion visualizer: The visual summary of consumers' basic biological and psychological emotions is presented by the emotion visualizer. It utilizes two types of visualizations: bar charts and line graphs to display emotions expressed in online reviews. A bar chart visualization is used to present the classification and summarization of consumers' biological and psychological emotions at product and feature levels. To display the temporal and volatility of reviews, a line graph visualization is utilized by the emotion visualizer. It takes the file containing features, emotion words, and emotion categories from the emotion classifier and aggregates these emotions according to the emotion categories at the product and feature levels. Then, it plots the temporal dynamism and volatility of emotions using a line graph.

IV. Experimental Dataset, Results, and Discussion

This section presents the details of the experimental dataset, the results of experiments performed on the experimental dataset, and a discussion of the results.

A. Experimental Dataset

The proposed system is evaluated on online reviews from amazon.com. This dataset was used by many studies, including [23][7][73][80]. The data set contains consumers' reviews of five diverse digital products: two digital cameras (Canon G3 and Nikon Coolpix 4300), one mobile phone (Nokia 6610), one MP3 player (Creative LabsNomad Jukebox Zen Xtra 40GB), and one DVD player (Apex AD2600 Progressive-scan DVD player). However, the date metadata of the reviews is not available in this dataset. The authors added the metadata to the dataset.

B. Experimental Results and Discussion

This section presents the aggregated consumers' emotions at product and feature levels, along with the change in the consumers' emotions over time.

1) Consumers' emotion categories at the product level: The results of fine-grained and binary emotion classifications of the Canon G3 camera at the product level are presented in this section. Fig. 3 shows the fine-grained consumers' basic emotions classification of the Canon G3 Camera. The joy emotion is in the first position, having a value of 157, as consumers expressed the Joy emotion more compared to other emotions. Joy is considered a positive emotion; therefore, it can be inferred that consumers enjoyed the use of the Canon G3 camera. The anger emotion is in second position with a value of 102 which is less than the value of joy emotion depicting the consumers' dissatisfaction. Consequently, consumers also expressed other negative emotions: sadness, fear, and disgust with the values of 89, 80, and 50, respectively. In contrast, some consumers defined Canon G3 with surprise emotion (81 value). It can be concluded from Fig. 3 that most consumers are happy with the camera as the frequency of Joy is very high compared to other emotions. Fig. 4 represents the binary emotion classification of the Canon G3 camera presenting that consumers are not satisfied with the camera as they expressed negative emotions (anger, fear, disgust, and sadness) more than positive emotions (joy and surprise).

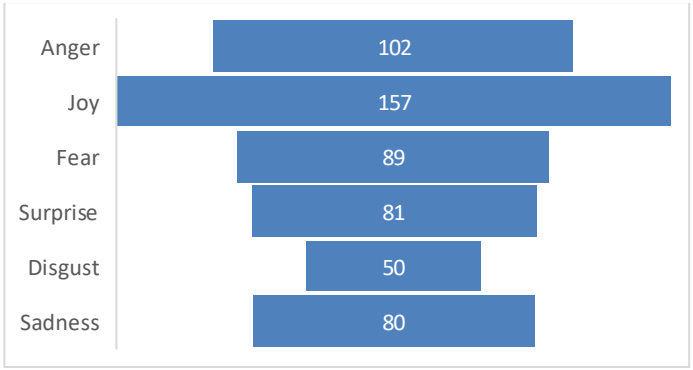


Fig. 3. Basic consumers' emotions on the Canon G3 camera.

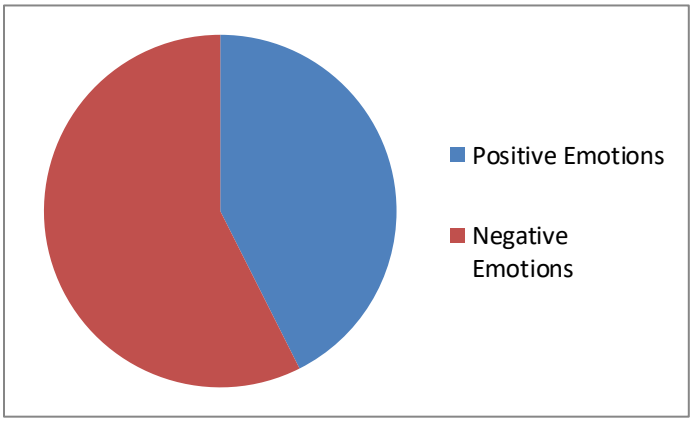


Fig. 4. Binary emotion classification of Canon G3 Camera.

2) Consumers' emotion categories at feature level: This section presents the basic consumers' emotions at the feature level. Due to the space limitations, we are presenting the fine-grained emotions classification of the top three features

(Camera, Picture, and Use) of the Canon G3 Camera (Fig. 5). Consumers expressed all emotions only through the camera feature. Disgust emotion is not mentioned by consumers for the picture and use features, highlighting that consumers like the picture and the use features of the camera. Moreover, only positive emotions (joy and surprise) are described by consumers for the use feature, showing the appreciation by the consumers towards the use of the camera. Further, it can be concluded that the camera is easy to use. All of these features are discussed using joy emotion remarkably as compared to other emotions. In other words, consumers are satisfied with these features tremendously. Specifically, for the camera, the highest discussed emotion is joy, with the highest value of 23 showing the endorsement of the camera by consumers. A few consumers are angry about the camera as shown by the value (4) of angry emotion. Also, other emotions (fear, surprise, disgust and sadness) are discussed by fewer consumers as indicated by a value of 3 for all features. Similarly, consumers exhibited mixed emotions (anger, surprise, and sadness) for the picture feature in a very small frequency of 3.

The binary emotion classification of the top three features is depicted in Fig. 6. All of these features received more positive emotions than negative emotions. The use feature received only positive emotions.

3) Temporal dynamism and volatility of consumers' emotions at the product level: This section describes the temporal dynamism and volatility of consumers' basic emotions are product level. Fig. 7 exhibited the temporal dynamism and volatility of consumers' emotions at product level. Consumers expressed all basic emotions from Q1 to Q4. More emotions are expressed in Q1 as shown by the values of anger, joy, fear, surprise, disgust, and sadness. On the other hand, fewer emotions are expressed in other quarters of the year. Fig. 8 shows the temporal dynamism and volatility of consumers' emotions at product level over a period of one year. In the month of March, the joy emotions are expressed more by the consumers. It is also interesting to note that at the starting month, consumers described the Canon G3 camera with joy and emotion.

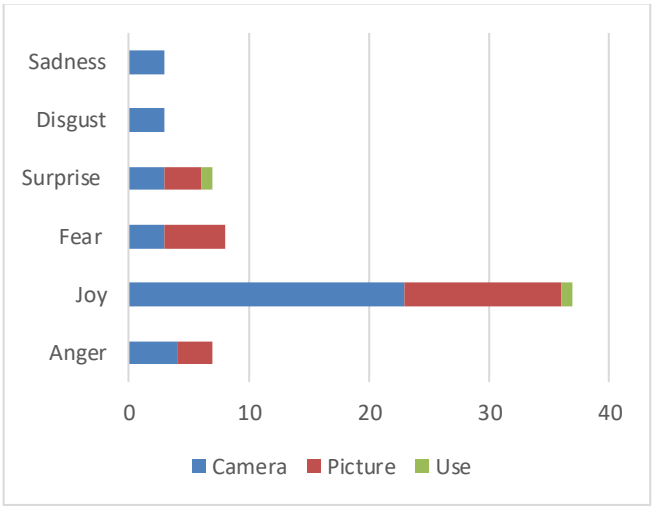


Fig. 5. Basic consumers' emotions at the feature level.

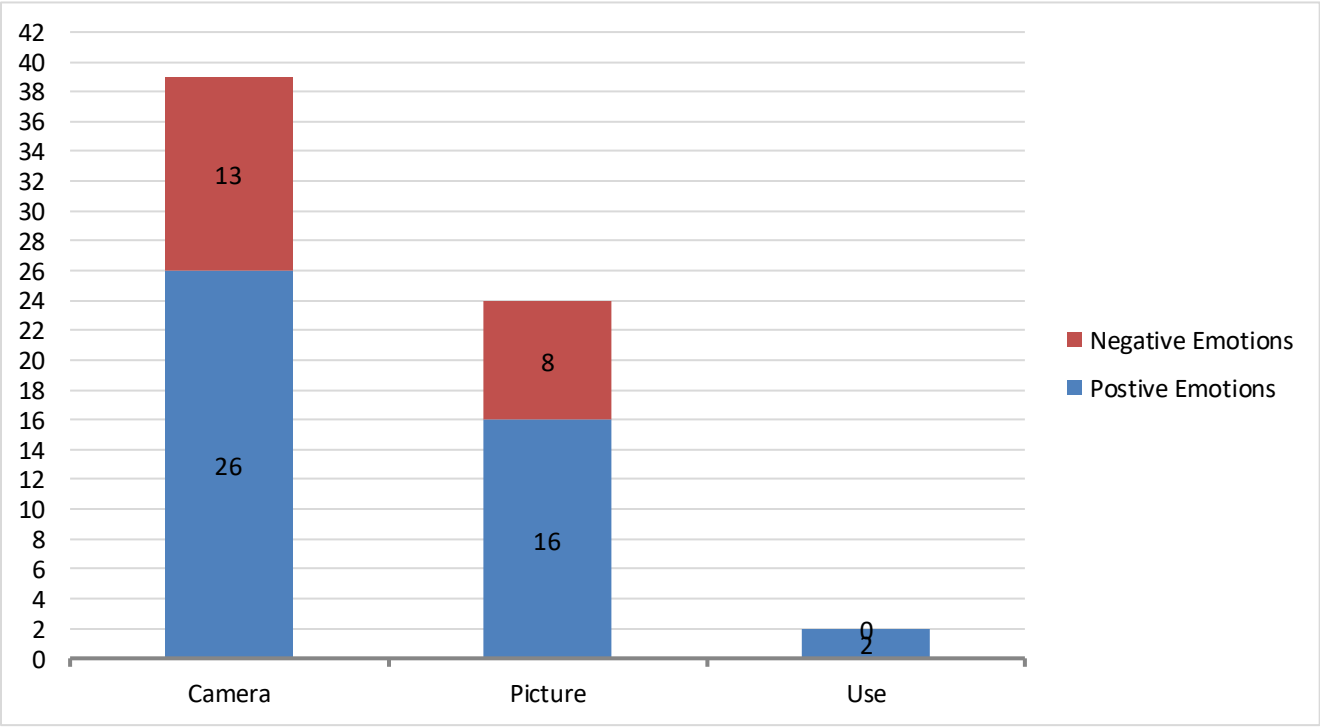


Fig. 6. Binary classification of the top three features.

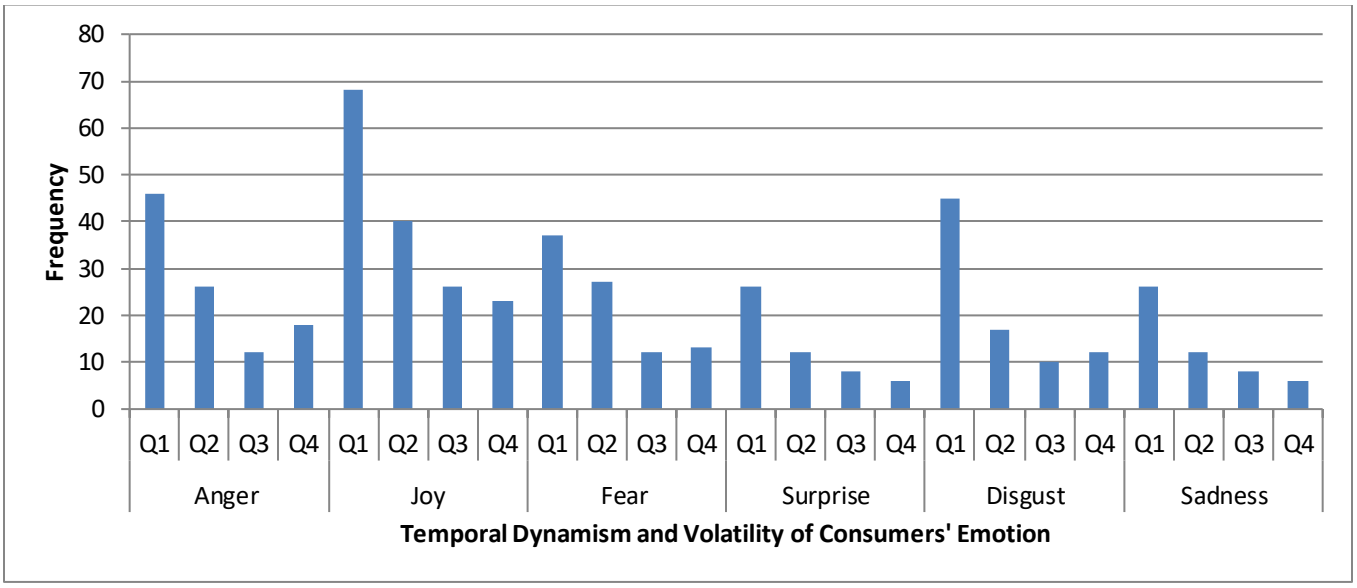


Fig. 7. Temporal dynamism and volatility of basic consumers' emotions at the product level.

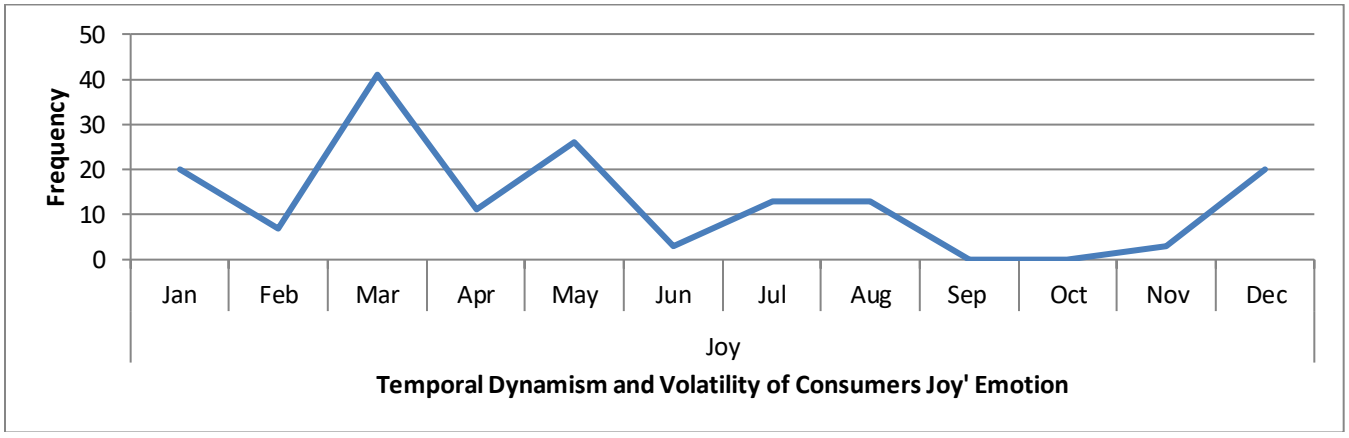


Fig. 8. Temporal dynamism and volatility of joy emotion at the product level.

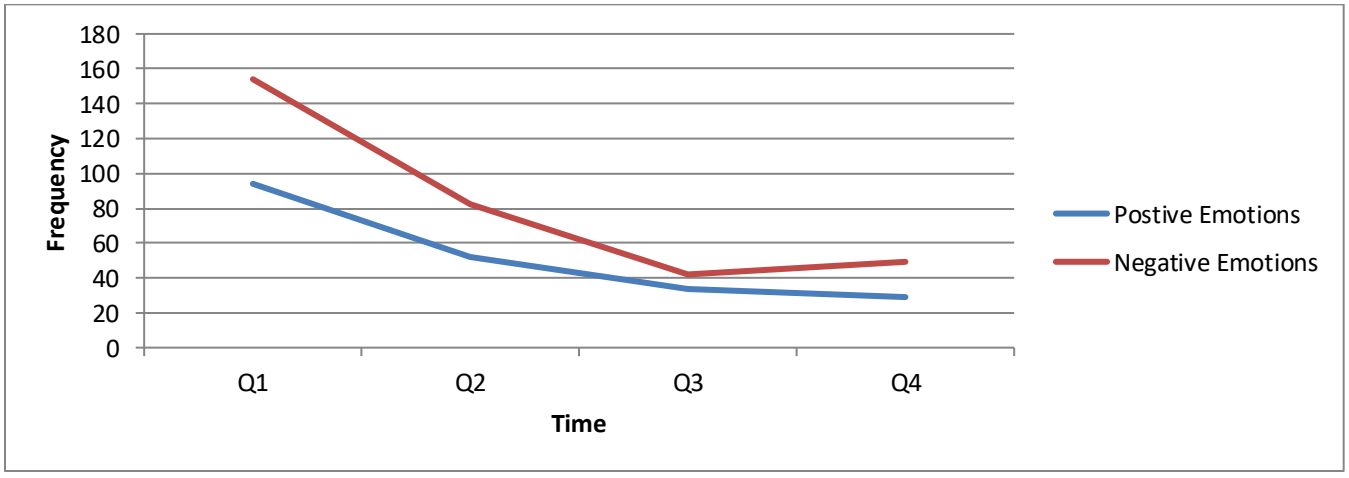


Fig. 9. Temporal dynamism and volatility of binary consumers' emotions at the product level.

Fig. 9 describes the temporal dynamism and volatility of binary consumers' emotions at the product level. It is important to note that more negative emotions are expressed in all quarters of the year as compared to positive emotions. Therefore, it shows the consumers' dissatisfaction with the camera. There is a decreasing trend in both positive and negative emotions

expression. However, in Q1 and Q2, consumers are more expressive than in the other quarters of the year (Q3 and Q4).

4) Temporal dynamism and volatility of consumers' emotions at the feature level: Fig. 10 exhibited the temporal dynamism and volatility of consumers' emotions at the feature level. It can be seen from the figure that the picture feature is discussed only in Q1. Therefore, it can be concluded that with the launch of the camera, the consumers expressed more emotions on prominent features like picture, battery, etc. However, with time, the focus of consumers' emotions is shifted to some less critical features.

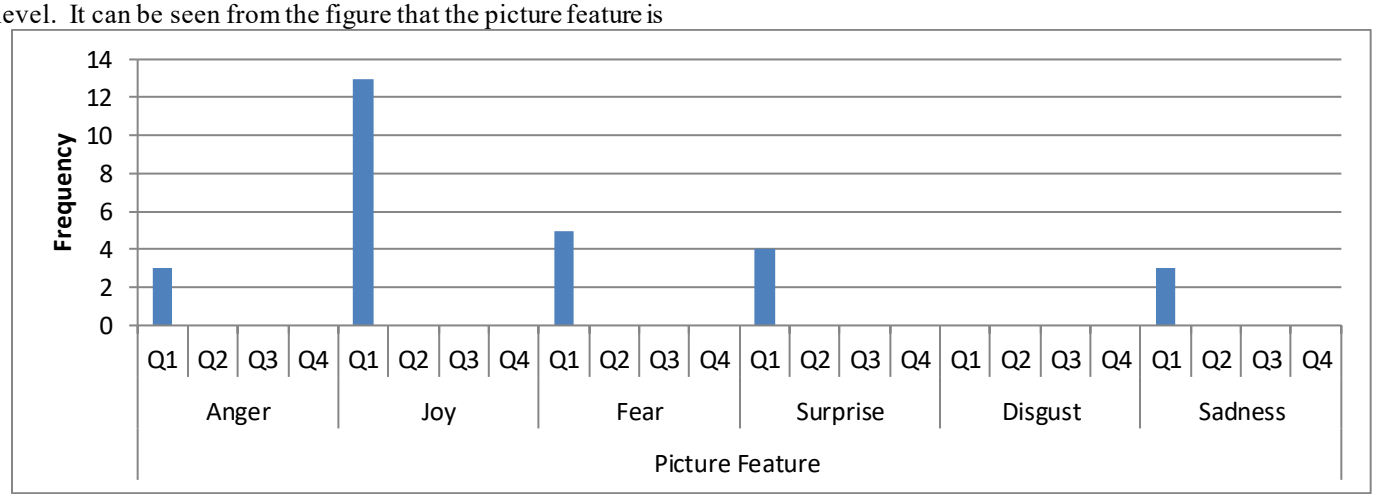


Fig. 10. Temporal dynamism and volatility of basic consumers' emotions at the feature level.

V. Research Implications

A. Business Perspective

The implications of expected results from a business perspective are for corporate decision-makers. First, it aids corporate decision-makers to improve their products and services based on the negative EWOM consumers. Second, the critical product features (the features having a great impact on sales) can be identified and can be targeted in future decision-making. Thirdly, the critical identified features can be focused on in new product development to enhance future sales. Fourthly, the result employing positive EWOM can be utilized for the reputation management of the organization to attract more consumers. Further, new product advertisements having positive feedback from consumers may result in increased consumer loyalty and trust, along with new consumers. Lastly, the marketing strategies can be designed based on expected results to attain consumers' attention and retention.

B. Consumers' Perspective

From the consumers' perspective, the results can be exploited in two different ways. First, the results will highlight the relative strengths and weaknesses of the target product (s) on a deeper (fine-grained) level that facilitates consumers' decision-making process. Moreover, the outcomes will depict the true emotions, evaluations, opinions, and judgments of consumers without any bias, as the EWOM posed no constraints on free expression. Consequently, the outcomes support consumers in their purchase and product adoption decisions.

VI. Conclusion, Limitations, and Future Directions

EWOM is remarkably and consistently increasing due to the active participation of consumers over different online forums, including online reviews. However, EWOM is unable to present consumers' emotions on different products and services. Specifically, basic biological and psychological emotions need to be extracted, summarized, and visualized from EWOM to empower consumers and enterprises to make informed decisions. Therefore, this research presents emotion categories and the temporal dynamism of emotion expressed in online reviews. The study extracted, summarized, and presented six basic emotion categories: anger, fear, disgust, joy, surprise, and sadness of well-known emotion model. Joy emotion is expressed more by the consumers in the online reviews of the Canon G3 camera published on amazon.com, indicating the satisfaction of the consumers towards the camera. A few consumers criticize the picture feature of the camera by describing it with negative emotions (sadness, anger, and fear). The proposed system has some potential limitations. First, it is evaluated on 332 reviews of electronic products. Second, the reviews are from amazon.com. The future direction of work includes the analysis of the proposed system by using a large number of reviews. Further, the proposed system can be utilized to analyze online reviews from different domains, like movie reviews and hotel reviews.

Funding

This work was funded by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, under grant No. UJ-02-074-DR. The authors, therefore, acknowledge with thanks DSR's technical and financial support.

References

[1] O'Brien, C., The emergence of media empowered consumer. Irish marketing review, 2011. 21(1/2): p. 32-40.

[2] Hennig-Thurau, T., et al., Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? Journal of interactive marketing, 2004. 18(1): p. 38-52.

[3] Na, J.C., T.T. Thet, and C.S. Khoo, Comparing sentiment expression in movie reviews from four online genres. Online Information Review, 2010.

[4] Chatterjee, S., Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. International Journal of Hospitality Management, 2020. 85: p. 102356.

[5] Fuad, Ammar, and Aida Fitri. "THE INFLUENCE OF DIGITAL MARKETING AND ELECTRONIC WORD OF MOUTH (E-WOM) ON PURCHASE INTENTION." *Global Research in Economics and Advanced Theory (GREAT)* 3.2 (2026): 91-101.

[6] Anastasiei, Bogdan, Nicoleta Dospinescu, and Octavian Dospinescu. "Individual and product-related antecedents of electronic word-of-mouth." *Aslib Journal of Information Management* 78.2 (2026): 449-469.

[7] Shamim, A., et al., Critical product features' identification using an opinion analyzer. The Scientific World Journal, 2014. 2014.

[8] Wang, Shan, et al. "Content dissimilarity and online review helpfulness: Contextual insights." *Journal of Business Research* 187 (2025): 115068.

[9] ADENIRAN, Adetayo Olaniyi, Goodness OLALEYE, and Ayobami Abdulmateen GBADEGESIN. "Influence of Online Reviews on Customer Satisfaction in E-Commerce Logistics Firm: Evidence from TEMU Nigeria." *Journal International Review of Research Studies* 1.04 (2026): 1-27.

[10] Singh, Amit, et al. "Quantifying the effect of eWOM embedded consumer perceptions on sales: An integrated aspect-level sentiment analysis and panel data modeling approach." *Journal of Business Research* 138 (2022): 52-64.

[11] Hao, M., et al. Visual sentiment analysis on twitter data streams. in 2011 IEEE Conference on Visual Analytics Science and Technology (VAST). 2011: IEEE.

[12] Li, H.; Bruce, X.B.; Li, G.; Gao, H. Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. Tour. Manag. 2023, 96, 104707.

[13] Ray, Rajeev Kumar, and Amit Singh. "From online reviews to smartwatch recommendation: An integrated aspect-based sentiment analysis framework." *Journal of Retailing and Consumer Services* 82 (2025): 104059.

[14] Abbasi Moghaddam, S., Aspect-based opinion mining in online reviews. 2013, Applied Sciences: School of Computing Science.

[15] Chen, C.; Zhang, D.; Zhang, F.; Zhang, F. The interaction effect of online composite reviews and time pressure on information persuasiveness: A heuristic-systematic perspective. Electron. Commer. Res. Appl. 2023, 59, 101269.

[16] Lin, Y.; Jen, L. The impact of consumer heterogeneity in the product life cycle on the diffusion patterns of user reviews and sales. J. Retail. Consum. Serv. 2024, 76, 103558.

[17] Eirinaki, M., S. Pisal, and J. Singh, Feature-based opinion mining and ranking. Journal of Computer and System Sciences, 2012. 78(4): p. 1175-1184.

[18] Ahmad, T. and M.N. Doja, Ranking system for opinion mining of features from review documents. International Journal of Computer Science Issues (IJCSI), 2012. 9(4): p. 440.

[19] Zhang, L., et al. Extracting and ranking product features in opinion documents. in Coling 2010: Posters. 2010.

[20] Zhang, K., R. Narayanan, and A.N. Choudhary, Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking. WOSN, 2010. 10: p. 11-11.

[21] Gu, S., et al., A model for basic emotions using observations of behavior in Drosophila. Frontiers in psychology, 2019. 10: p. 781.

[22] Khan, S.S., et al., CHALLENGES IN OPINION MINING, COMPREHENSIVE.

[23] Shamim, A., et al., Multi-Attribute Online Decision-Making Driven by Opinion Mining. Mathematics, 2021. 9(8): p. 833.

[24] Ullah, R., et al., From valence to emotions: Exploring the distribution of emotions in online product reviews. Decision support systems, 2016. 81: p. 41-53.

[25] Ekman, P., An argument for basic emotions. Cognition & emotion, 1992. 6(3-4): p. 169-200.

[26] Bruna, O., H. Avetisyan, and J. Holub. Emotion models for textual emotion classification. in Journal of Physics: Conference Series. 2016: IOP Publishing.

[27] McDougall, William, "The hypothesis of inhibition by drainage." *Psychological Review* 33.5 (1926): 370.

[28] Watson, John Broadus. "Behaviorism, rev." (1930).

[29] Arnold, Magda B. "Emotion and personality." (1960).

[30] Mowrer, Orval. "Learning theory and behavior." (1960).

[31] Tomkins, Silvan S. "Surprise—startle: The resetting affect." (1962).

[32] Izard, Carroll E. "Differential emotions theory." *Human emotions*. Boston, MA: Springer US, 1977.

[33] Plutchik, Robert. "A general psycho evolutionary theory of emotion." *Theories of emotion*. Academic press, 1980. 3-33. 43-66.

[34] Panksepp, Jaak. "Archaeology of mind." *Behavioral and Brain Sciences* 5.3 (1982): 449-467.

[35] Weiner, Bernard, and Sandra Graham. "An attributional approach to emotional development." *Emotions, cognition, and behavior* (1984): 167-191.

[36] Frijda, Nico H. "The emotions." *Studies in Emotion and Social Interaction* (1986).

[37] Shaver, Phillip, et al. "Emotion knowledge: further exploration of a prototype approach." *Journal of personality and social psychology* 52.6 (1987): 1061.

[38] Richins, Marsha L. "Measuring emotions in the consumption experience." *Journal of consumer research* 24.2 (1997): 127-146.

[39] Derbaix, Christian, and Joëlle Vanhamme. "Inducing word-of-mouth by eliciting surprise–a pilot investigation." *Journal of economic psychology* 24.1 (2003): 99-116.

[40] Ekman, Paul. "Are there basic emotions?." (1992): 550.

[41] Yadollahi, A., A.G. Shahraki, and O.R. Zaiane, Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 2017. 50(2): p. 1-33.

[42] Amini, R., C. Lisetti, and G. Ruiz, HapFACS 3.0: FACS-based facial expression generator for 3D speaking virtual characters. IEEE Transactions on Affective Computing, 2015. 6(4): p. 348-360.

[43] Teijeiro-Mosquera, L., et al., What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. IEEE Transactions on Affective Computing, 2014. 6(2): p. 193-205.

[44] Balog, K., G. Mishne, and M. Rijke. Rijke. why are they excited? identifying and explaining spikes in blog mood levels. in In Proceedings 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL. 2006)

[45] Pennebaker, J.W., Writing about emotional experiences as a therapeutic process. Psychological science, 1997. 8(3): p. 162-166.

[46] Zamani, N.A.M., et al. Sentiment analysis: determining people's emotions in Facebook. in Proceedings of the 13th international conference on applied computer and applied computational science. 2014.

[47] Mohammad, S.M., et al., Sentiment, emotion, purpose, and style in electoral tweets. Information Processing & Management, 2015. 51(4): p. 480-499.

[48] Bhowmick, P.K., A. Basu, and P. Mitra, Reader Perspective Emotion Analysis in Text through Ensemble based Multi-Label Classification Framework. Computer and Information Science, 2009. 2(4): p. 64-74.

[49] Lei, J., et al., Towards building a social emotion detection system for online news. Future Generation Computer Systems, 2014. 37: p. 438-448.

[50] Gupta, N., M. Gilbert, and G.D. Fabbrizio, Emotion detection in email customer care. Computational Intelligence, 2013. 29(3): p. 489-505.

[51] Luo, B., J. Zeng, and J. Duan, Emotion space model for classifying opinions in stock message board. Expert Systems with Applications, 2016. 44: p. 138-146.

[52] Min, H., et al., Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. Journal of Applied Psychology, 2021. 106(2): p. 214.

[53] Adikari, A., et al., Emotions of COVID-19: Content Analysis of Self-Reported Information Using Artificial Intelligence. Journal of Medical Internet Research, 2021. 23(4): p. e27341.

[54] Plutchik, R., A psycho evolutionary theory of emotions. 1982, Sage Publications.

[55] Bollen, J., H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. in Proceedings of the international AAAI conference on web and social media. 2011.

[56] Wang, W., et al. Harnessing twitter big data for automatic emotion identification. in International Conference on Privacy, Security, Risk and Trust 2012

[57] Hasan, M., E. Agu, and E. Rundensteiner. Using hashtags as labels for supervised learning of emotions in twitter messages. in ACM SIGKDD workshop on health informatics, New York, USA. 2014.

[58] Mishne, G. Experiments with mood classification in blog posts. in Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access. 2005.

[59] Mishne, G. and M. De Rijke. Capturing Global Mood Levels using Blog Posts. in AAAI spring symposium: computational approaches to analyzing weblogs. 2006.

[60] Alm, C.O., D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. in Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005.

[61] Luyckx, K., et al., Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. Biomedical informatics insights, 2012. 5: p. BII. S8966.

[62] Jain, M.C. and V. Kulkarni, TexEmo: Conveying emotion from text-the study. International Journal of Computer Applications, 2014. 86(4).

[63] Ho, D.T. and T.H. Cao. A high-order hidden Markov model for emotion detection from textual data. in Pacific Rim Knowledge Acquisition Workshop. 2012: Springer.

[64] Zablocki, A., K. Makri, and M.J. Houston, Emotions within online reviews and their influence on product attitudes in Austria, USA and Thailand. Journal of Interactive Marketing, 2019. 46: p. 20-39.

[65] Mokryn, O., et al., Sharing emotions: determining films' evoked emotional experience from their online reviews. Information Retrieval Journal, 2020. 23(5).

[66] Garcia, D. and F. Schweitzer. Emotions in product reviews--empirics and models. in 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. 2011: IEEE.

[67] Yin, G., Q. Zhang, and Y. Li, Effects of emotional valence and arousal on consumer perceptions of online review helpfulness. 2014.

[68] Moghaddam, S. and M. Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. in Proceedings of the 19th ACM international conference on Information and knowledge management. 2010

[69] Tanase, D., et al., Emotions and activity profiles of influential users in product reviews communities. Frontiers in Physics, 2015. 3: p. 87.

[70] Hossain, M.S.; Rahman, M.F. Customer sentiment analysis and prediction of insurance products' reviews using machine learning approaches. FIIB Bus. Rev. 2023, 12, 386–402.

[71] Kim, J. and P. Gupta, Emotional expressions in online user reviews: How they influence consumers' product evaluations. Journal of Business Research, 2012. 65(7): p. 985-992.

[72] Luo, Y.; Li, Y.; Ye, Q. Impacts of online additional reviews on the sales volume of cross-border pharmaceutical e-commerce platforms. J. Glob. Inf. Technol. Manag. 2022, 25, 83–101.

[73] Yang, J.-Y., H.-j. Kim, and S.-g. Lee, Feature-based Product Review Summarization Utilizing User Score. J. Inf. Sci. Eng., 2010. 26(6): p. 1973-1990.

[74] Castro, F., et al., Applying data mining techniques to e-learning problems, in Evolution of teaching and learning paradigms in intelligent environment. 2007, Springer. p. 183-221.

[75] Arnold, K.E., Z. Tanes, and A.S. King, Administrative perceptions of data-mining software Signals: Promoting student success and retention. The Journal of Academic Administration in Higher Education, 2010. 6(2): p. 29-39.

[76] Mohammad, Saif M., and Peter D. Turney. "Crowdsourcing a word–emotion association lexicon." *Computational intelligence* 29.3 (2013): 436-465

[77] Ali, L., et al., Factors influencing beliefs for adoption of a learning analytics tool: An empirical study. Computers & Education, 2013. 62: p. 130-148.

[78] Das, S. and A. Dutta, Characterizing public emotions and sentiments in COVID-19 environment: A case study of India. Journal of Human Behavior in the Social Environment, 2021. 31(1-4): p. 154-167.

[79] Wen, Zhanming, et al. "Text Mining Based Approach for Customer Sentiment and Product Competitiveness Using Composite Online Review Data."*Journal of Theoretical and Applied Electronic Commerce Research* 19.3 (2024): 1776-1792.

[80] Ruskanda, Fariska Zakhralativa, Dwi Hendratmo Widyantoro, and Ayu Purwarianti. "Efficient Utilization of Dependency Pattern and Sequential Covering for Aspect Extraction Rule Learning." *J. ICT Res. Appl* 14.1 (2020): 51.

Privacy Preserving Federated Graph Learning with Data Envelopment Analysis Driven Interpretable Customer Segmentation Framework

Pentareddy Ashalatha[1], Dr. G. Krishna Mohan[2]
Research Scholar, Department of Computer Science & Applications, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[1]
Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[2]

Abstract*—Customer segmentation plays a critical role in retail analytics by enabling personalized marketing, optimized resource allocation, and data-driven strategic decision-making. However, customer data is often distributed across multiple retail branches and contains sensitive transactional information, creating significant challenges related to privacy preservation, regulatory compliance, and model interpretability. Traditional segmentation approaches, including clustering algorithms and centralized deep learning methods, typically require aggregated data storage, which increases privacy risks and limits secure deployment in distributed retail environments. In addition, many existing methods fail to capture complex relational patterns such as co-purchasing behavior and inter-customer structural dependencies. To address these limitations, this study proposes FedGraph-DEA, a novel hybrid framework integrating federated learning, graph neural networks (GNNs), and Data Envelopment Analysis (DEA) for privacy-preserving and efficiency-aware customer segmentation. The framework first employs Distributed Federated Convolutional Autoencoders to extract latent customer representations from decentralized retail datasets. Similarity graphs are then constructed locally, followed by federated GNN-based community detection to identify structurally coherent customer groups without sharing raw data. Finally, DEA is applied to evaluate the operational efficiency of the discovered customer segments. Experimental evaluation was conducted using the UCI Online Retail dataset partitioned across five simulated non-IID client nodes. The proposed model achieved 96.1% accuracy, 0.95 precision, 0.96 recall, and 0.95 F1-score when compared with pseudo-ground-truth clusters generated through K-Means reference clustering. Furthermore, the framework obtained a silhouette score of 0.74 and a modularity value of 0.57, demonstrating strong cluster compactness and structural separation. The proposed system provides a scalable, interpretable, and privacy-preserving solution for distributed retail analytics.

Keywords*—Customer segmentation; federated learning; graph neural networks; data envelopment analysis; privacy preservation

I. Introduction

In today's digital economy, customer segmentation is generally recognized as a pillar of data-driven decision-making and marketing [1], [2]. By segmenting customers into substantive groups, companies can tailor services to individuals, improve resource usage, and create targeted campaigns [3]. Nevertheless, with the exponential growth of online transactions and widespread sensitive customer data proliferation [4], privacy-enhancing and explainable methods for segmentation have become ever more important [5]. Regulatory protocols like GDPR and HIPAA mandate rigorous adherence to privacy principles, as businesses at the same time seek transparent and actionable information from their analytics platforms [6]. Conventional centralized approaches leave organizations vulnerable to data leakage, breaches, and regulatory non-compliance, and thus, there is a need for paradigms of federated learning [7]. Of equal significance is interpretability, as stakeholders and managers must not only know what segments exist but also the reason why those segments are effective and worthwhile [8]. In this context, privacy-preserving interpretable customer segmentation has become a salient area of attention in academia and the industry alike, filling the gap between secure distributed learning and decision-driven insights [9].

Previous approaches to customer segmentation have primarily utilized statistical clustering algorithms like K-Means, DBSCAN [10], [11], and hierarchical clustering, or machine learning classifiers like Random Forests and XGBoost [12], [13]. Although these approaches provide computational ease, the models tend to lack flexibility in accommodating high-dimensional transactional data and are not robust in distributed settings [14]. In addition, the model tends to need centralized data collation, which presents privacy threats in areas like finance, healthcare, and e-commerce [15]. Autoencoders based on deep learning have also recently been considered for extracting latent behavior features; however, without federated approaches, the models still need to pool raw data at a central server[16]. In terms of interpretability, clustering and classification methods usually return black-box outputs, providing decision-makers with little understanding of the efficiency or strategic importance of the segmented findings [17]. Therefore, current frameworks fail to meet the two-pronged challenge: protecting sensitive customer information in decentralized ecosystems and returning interpretable, transparent results for managers [18], [19].

The FedGraph-DEA model plan creates a customer segmentation framework that is privacy-protected and interpretable, and is designed to work in a distributed retail-

setting. DFCA is used to learn high-quality latent behavioral representations using transactional data that is decentralized without the exchange of raw information. These similarity embeddings are organized into customer similarity graphs in order to obtain co-purchase dependencies and multi-hop relational patterns. With the help of Federated Graph Neural Networks (GNNs), the recognition of coherent communities of customers in non-IID data distributions is then realized. Lastly, Data Envelopment Analysis (DEA) measures the operational efficiency of every segment, which allows for actionable performance-based and interpretable segmentation results by managers.

A. Research Motivation

The rise of the distributed retail systems has escalated the necessity of customer segmentation strategies that adequately protect data privacy and yet express intricate behavioral relationships [20]. Conventional centralized systems reveal sensitive customer data and do not represent relational relationships like co-purchase behavior. Simultaneously, the existing models are usually not interpretable, which constrains their applicability in managerial decision-making. The rationale behind this research is to come up with a single framework, which is able to take care of privacy preservation, relational intelligence, and transparency, so as to institute the correct and practical customer segmentation in the retail settings where data is sensitive and decentralized.

B. Research Significance

This research is significant because it attempts to fill vital shortcomings of current customer segmentation methods by combining the concepts of privacy protection, relational modelling, and interpretability under a single system. Contrary to the old-fashioned approaches, it provides a secure data processing in distributed contexts, as well as, the transformation of intricate customer interactions and offers efficiency-related insights to make a decision. The research has also helped the academic research as well as practical applications through facilitating scalable, transparent, and performance-driven analytics. It can be applied to privacy sensitive areas, like store, finance and healthcare where data analysis requires the use of secure and readable data.

C. Problem Statement

There are several limitations associated with the current techniques of customer segmentation in a distributed retail setting. Firstly, current traditional approaches of customer segmentation tend to aggregate private and sensitive information about customers, exposing the system to higher risks of data leaks and non-conformity with relevant regulations [21]. Finally, the majority of current models used for customer segmentation have no interpretability; consequently, they provide little information about the factors that affect customer segmentation [22]. In spite of the attempts by modern researchers to introduce federated learning models to address privacy and relational intelligence challenges [23], interpretability and efficiency-based assessment of customer segments remain underestimated aspects. Thus, the development of a unified approach to customer segmentation seems to be essential.

D. Key Contribution

Introduced a customer segmentation algorithm that maintains customer privacy to solve the problem of data confidentiality, scalability, and distributed retail environment constraints as well as performs behavioral representation learning without any data aggregation in the centralized setting.

Provides the use of DEA as an additional post-segmentation evaluation tool to measure the efficiency of the graph-based customer communities formed based on operational efficiency.

Proposes a privacy-preserving federated similarity graph formation mechanism to capture multi-hop dependency among the customers' behavior patterns in non-IID distributed learning environment without sharing raw customer data among the retailers.

Combines Federated Graph Neural Network methods with Distributed Federated Convolutional Autoencoder (DFCA), allowing joint modeling of compressed behaviors and relations in the decentralized retail setting demonstrates that the combined DFCA + Federated GNN + DEA framework improves segmentation quality, clustering coherence, and interpretability compared to conventional centralized and federated baseline models.

The research is presented in the following structure: Existing studies are presented in Section II. Section III covers problems with the existing study. The suggested methodology is presented in Section IV. Section V includes the results and a discussion of this research. Conclusions and future works are covered in Section VI.

II. Related Work

Le et al. in [24] presents PersonalFR, a personalized federated recommender system that utilizes autoencoder-based models to learn user-item interactions without revealing raw data. The local updating of encoders and aggregation of decoders on the server side maintains data privacy. The experiments on real-world datasets show that PersonalFR offers performance close to centralized models and with lower communication overhead. The non-interpretability of the autoencoder's latent features also restricts the explainability of the segmentation outcomes. J. M. John, O. Shobayo, and B. Ogunleye in [25] compares different clustering algorithms—K-means, Gaussian Mixture Models (GMM), DBSCAN, Agglomerative Clustering, and BIRCH—on a UK-based online retail dataset with more than 500,000 records. Based on the RFM (Recency, Frequency, and Monetary) framework, the study seeks to improve decision-making in the retail industry. Nonetheless, the use of centralized data processing in the study creates data privacy and scalability issues. M. Alves Gomes and T. Meisen in [26] evaluates the 105 publications published between 2000 and 2022. The four stages of the process that are identified in the work are data collection, customer representation, segmentation, and targeting. The integration of sophisticated approaches such as PCA and SOMs is also described in the review.

Wang in [27] investigates the coupling of deep learning models, customer segmentation, online marketing, and swarm

intelligence algorithm models. The synergy aims at maximizing the performance of the clustering by the global search capability of swarm intelligence. The study does not address the problem of data heterogeneity in distributed environments. In addition, the interpretability of the resulting customer segments is constrained, as well as integration with efficiency assessment frameworks.

A. Sharma, N. Patel, and R. Gupta in [28] investigates the convergence of deep learning methods with predictive analytics to enhance customer segmentation techniques. Through the use of neural networks to examine intricate customer data, the study seeks to reveal subtle behavioral patterns that may be missed by conventional methods. Nevertheless, the study admits difficulties in model interpretability and the necessity for large computational resources, which could restrict its usage for smaller businesses or companies with limited technical infrastructure. Y. Long et al.,[29] Presented a federated deep clustering method for retail customer segmentation, integrating local autoencoders with a central K-means clustering model. Using federated learning, their model ensured customer data privacy across different branches while maintaining an overall segmentation accuracy of around 88%. But the study was not concerned with the evaluation of the resulting clusters in terms of operational efficiency or business value, which curtails its applicability in strategic decision-making. Contrarily, the current FCA–DEA framework not only improves privacy via federated training but also encompasses a DEA-based layer to evaluate and rank customer segments, which fills the gap between high-performance clustering and business interpretability in real-world scenarios.

P. Hu et al., [30] proposed a federated CNNs-based privacy-aware clustering model to partition customer data scattered over multiple retail stores. Their approach effectively circumvented raw data centralization and had a clustering accuracy of approximately 90%, which proved the feasibility of federated deep learning for customer analytics. However, the approach did not incorporate an interpretability function to justify the relevance of the clusters and provided no facility for benchmarking the operating efficiency of every segment. This limitation justifies the necessity for hybrid models such as the FCA–DEA framework, which pairs federated unsupervised learning with a post-segmentation DEA module for clear and actionable efficiency assessment.

Y. Jiang et al., [31] introduced federated VAE to use in distributed customers profiling in the retail industry. At their current stage their approach employed VAEs in order to elicit non-linear latent representations of customer purchasing tendency subject to privacy-respecting constraints with a reconstruction precision of approximately 92.1%. Although the model has shown robust features learning without focusing sensitive data, still, it did not consider the problem of making sense of the latent features to apply business results, as well as assessing the relative efficiency of the generated groups of customers. In comparison, the suggested FCA-DEA framework integrates deep representation learning with operational efficiency ranking of DEA so that businesses could translate latent features into plain, usable forms in terms of resource optimization and targeted marketing.

Based on the reviewed literature, it can be seen that past research either concentrated on centralized clustering and deep learning processes, which ignore privacy, or federated models, which do not analyze interpretability or operational efficiency. The graph-based model, though effective in depicting feature interaction still uses centralized access to data. The FedGraph-DEA framework will overcome such shortcomings by combining federated autoencoders to learn decentralized representations, graph neural networks to detect communities, and data envelopment analysis to identify the efficiency of the model. This is a hybrid design that provides secure learning, increased segmentation accuracy, and transparent decision support, something that previous models did not have.

III. FedGraph-DEA Framework for Customer Segmentation

The FedGraph-DEA method employs the concept of a single pipeline that combines decentralization representation learning, graph-based feature interaction model, and efficiency-based interpretability. Every client node separately trains a Distributed Federated Convolutional Autoencoder (DFCA) to learn the customer latent embeddings without having access to raw data. These embeddings are then used to create a customer similarity graph based on a cosine affinity. The graph establishes the connectivity of the nodes and is enhanced using federated Graph Neural Network (GNN), which reflects on multi hop customer relations through federated weight aggregation across distributed subgraphs. Communities that are formed as a result are customer segments. These segments are then measured by Data Envelopment Analysis (DEA) wherein each of the segments is considered as a decision-making unit and the evaluation is done in terms of business efficiency by considering the marketing inputs and profit-related outputs. DFCA induced graph formation, GNN-based segmentation and DEA-based evaluation are new to interact. FedGraph-DEA is the only method where privacy preservation, relational intelligence, and interpretable efficiency assessment are integrated into a single framework in contrast to the current federated clustering, federated GNN, or single method of standalone DEA. Fig. 1 shows the overall workflow.

The innovation introduced by the FedGraph-DEA model is its comprehensive integration of three distinct yet complementary approaches in one privacy-preserving segmentation architecture for distributed retailers. In contrast to the traditional federated clustering models that concentrate exclusively on privacy-preserving capabilities, this new model is capable of analyzing higher-order customer relations with the aid of graph neural networks while concurrently measuring the efficiency of their operations with DEA.

The principal novelty of this methodology lies in the unified integration of federated representation learning, graph-based relational modeling, and efficiency-driven evaluation within a single decentralized analytical framework. Unlike conventional customer segmentation approaches that rely on centralized data aggregation and purely distance-based clustering, this architecture enables privacy-preserving latent feature extraction through federated convolutional autoencoders while simultaneously capturing higher-order behavioral dependencies via graph neural networks. The

transformation of learned embeddings into customer similarity graphs allows structural and multi-hop relationships to be incorporated into segmentation, addressing limitations of attribute-only models. Furthermore, the incorporation of Data Envelopment Analysis introduces an operational efficiency perspective that links discovered segments to measurable input–output performance metrics, enhancing managerial interpretability. This synergistic combination of privacy preservation, relational intelligence, and efficiency-based interpretability constitutes a novel contribution to distributed customer analytics research.

Fig. 1 illustrates the FedGraph-DEA approach for customer segmentation under the privacy-preserving setting. First, the UCI Online Retail dataset is considered, then the preprocessing phase involves data cleaning, normalization, and feature engineering. The next layers, known as federated learning, employ distributed autoencoder methods and secure aggregation to learn the latent space representation without sharing data. Using the embeddings, a graph structure can be built, which is utilized by federated graph neural networks for community detection. Finally, DEA is employed to analyze efficiency.

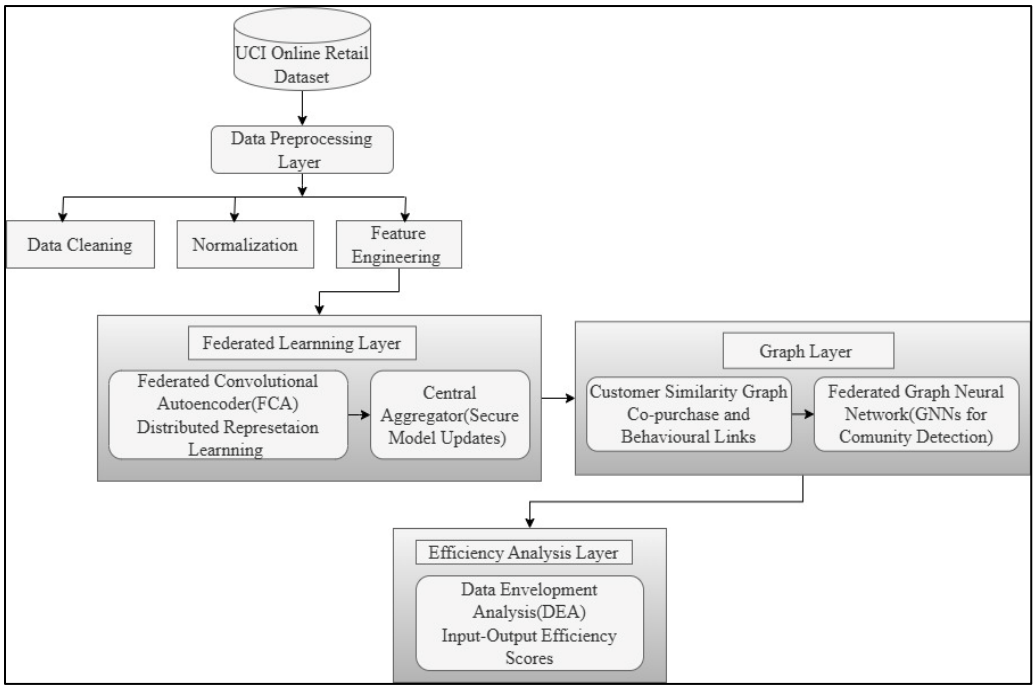


Fig. 1. Privacy-preserving federated graph-based customer segmentation workflow.

A. Data Collection

UCI Online Retail is an online retail dataset that entailed in excess of 500,000 transaction records in various nations in the 2010-2011 [32]. It stores specific buying behavior such as invoices, product characteristics, amount, costs, time values, customer names, and addresses. It is appropriate in terms of customer segmentation, federated learning, privacy-preserving analytics, and RFM behavioral modeling for distributed retail operations due to its heterogeneity and realistic nature. In the case of federated simulation, the database was split among five clients according to the country attribute for transactions, thus leading to naturally non-IID distributions, as buying habits and preferences vary greatly depending on location. This division allows the creation of distinct local databases for each client in an actual federated manner without the need for more than one data source physically.

Table I illustrates five sample transactions showing product codes, descriptions, quantities, and prices, demonstrating the dataset's structure and how individual purchase records form the basis for feature engineering and segmentation.

TABLE I. SAMPLE TRANSACTION RECORDS FROM THE UCI ONLINE RETAIL DATASET

Invoice No	Stock Code	Description	Quantity	Unit Price
536365	85123A	White Mug	6	2.55
536366	71053	Floral Notebook	3	4.95
536367	84406B	Pink Hanging Bag	2	6.95
536368	84029G	Heart Decoration	8	1.25
536369	22960	Jam Making Set	1	11.95

B. Data Preprocessing

Customer segmentation requires strict data preprocessing as it converts raw transactions into processed and meaningful patterns of customer behavior. This is essential in eliminating noise, ensuring data quality, and facilitating proper, reliable segmentation analysis.

1) Data cleaning: Missing Customer IDs will be removed and negative values will be deleted because that means returns or cancellations rather than purchases. Data integrity is ensured by discarding duplicate invoices. The resulting set of data only contains valid and applicable transactions.

2) Feature engineering: The highly set Recency-Frequency-Monetary (RFM) is used to extract customer-level behavioral attributes. The Eq. (1) and Eq. (2) define the calculations performed for each customer i,

$$R_i = t_{ref} - t_{last\ purch,i} \quad (1)$$

$$M_i = \sum_{j=1}^{n_i} p_{i,j} \quad (2)$$

where, R_i represents the period of time since the time of the last purchase made by the customer, F_i is the total number of purchases made, and M_i is the cumulative purchase value.

3) Normalization: Min-max normalization is used to rescales all numerical features, which ensure that the magnitude of the numerical feature is similar and scale is not dominant in federated and graph-based learning computed using Eq. (3):

$$X_{scaled} = \frac{.X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

where, X_scaled is the normalized feature value, X is the original data point, while X_{min} and X_{max} denote the minimum and maximum feature values used for scaling.

C. Distributed Federated Convolutional Autoencoder (DFCA)

The DFCA is the first module in the FedGraph-DEA framework and is used for training privacy-preserving latent embeddings for customers' transactional behaviors. Given that customer transactions are distributed over different branches and that their centralization is prohibited by privacy laws, each client will train an autoencoder using their normalized RFM feature vectors. Here, the encoder maps the feature vector into the embedded representation, while the decoder tries to reconstruct the feature vector from this latent embedding. The encoder working is determined as Eq. (4):

$$h_i^{(k)} = \sigma\left(W_{e,k} * x_i + b_{e,k}\right) \quad (4)$$

where $x_i \in \mathbb{R}^d$denotes the normalized RFM feature vector of customer i, $W_{e,k}$and $b_{e,k}$represent the encoder weight matrix and bias vector of client k, respectively, and $\sigma(\cdot)$ denotes a nonlinear activation function such as ReLU or sigmoid. The encoder is able to compress high-dimensional information about behavior into low-dimensional representation that is capable of preserving purchasing attributes despite noise and redundancy. The latent representation is subsequently reconstructed through the decoder network to preserve reconstruction fidelity, as expressed in Eq. (5):

$$\hat{x}_i^{(k)} = \sigma\left(W_{d,k} * h_i^{(k)} + b_{d,k}\right) \quad (5)$$

where $\hat{x}_i^{(k)}$ denotes the reconstructed customer feature vector, while $W_{d,k}$and $b_{d,k}$correspond to decoder parameters for client k. The decoder attempts to minimize information loss between the original and reconstructed customer behavior representations. The DFCA model is optimized using reconstruction loss, which measures the difference between the original and reconstructed feature vectors is given in Eq. (6),

$$\mathcal{L}_{\mathrm{rc}(k)}^{k} = \frac{1}{n_k}\sum_{i=1}^{n_k} |x_i - \hat{x}_i^{(k)}|_2^2 \quad (6)$$

where, N_k represents the number of customer samples available at client k . This objective ensures accurate preservation of local behavioral patterns while learning generalized latent embeddings suitable for federated graph construction and downstream segmentation tasks. In order to encourage the stability of models with highly non-IID purchasing distributions, and L_2 regularization factor can be defined as Eq. (7):

$$\mathcal{L}_{\mathrm{AE}}^{(k)} = (\mathcal{L}_{\mathrm{rc}}^{k} + \lambda|\theta_k|_2^2) \quad (7)$$

where, θ_k and ϕ_k represent decoder and encoder parameters, and λ controls the fine of huge weights. The $\theta_k = \{W_{e,k}, b_{e,k}, W_{d,k}, b_{d,k}\}$ refer to all the trainable parameters. The learning objective worked out in Eq. (8):

$$\min_{\theta_k} = \mathcal{L}_{AE}^{(k)} \quad (8)$$

where, θ denot $\mathcal{L}_{AE}^{(k)}$ represents its reconstruction loss; the objective minimizes this loss to learn optimal local latent representations from client data. This will guarantee uniform extraction of latent purchasing representations between different branches, stores or regions.

After local training, model parameters (but not raw data) are sent to an international aggregation server that is secure and is located centrally. The Federated Averaging (FedAvg) algorithm determines which client models are used in this process by means of a weighted average of the model whose dataset is the largest to the smallest. The international brief at the communication round $t+1$ is expressed can be defined as Eq. (9):

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n_{\mathrm{total}}} w_k^{(t)} \quad (9)$$

In this averaging process, $w_k^{(t)}$ are the local model weights of client k at the t th iteration. n_k is the size of the data sample of each node, and n total is the sum of all the samples. When aggregation has been done the server transmits the new global model to all the clients as received in Eq. (10):

$$\theta_k^{(t+1)} \leftarrow w^{(t+1)},\ n_{\mathrm{total}} = \sum_{k=1}^{K} n_k \quad (10)$$

where, $\theta_k^{(t+1)}$ denotes updated local parameters for client k, $w^{(t+1)}$ is the aggregated global model, n_k is client data size, and nn_{total} represents the total samples.

The synchronized update scheme allows converging to a common embedding space even in cases of statistical heterogeneity of distributed datasets. FedAvg is preferred because it is efficient in communication, does not suffer from instabilities in non-IID environments, and is scalable in industrial federated systems. Notably, the exchange of model updates only is provided, and the raw data do not leave their locations, which reduces the privacy leakage and ensures that the enterprise and regulatory data-protection policies are followed. The framework provides a uniform embedding space to clients through a combination of local representation learning with DFCA and global alignment with FedAvg. The

resulting embeddings are privacy-preserving, statistically resistant and behaviorally meaningful, which makes them a solid basis of building customer similarity graphs and makes them useful in graph-based community detection in the FedGraph-DEA architecture.

D. Customer Similarity Graph Construction

Once latent vectors are created by DFCA, the step that follows is to create a customer similarity graph that reflects behavioral relations among customers. Customers are represented as node and the relationships are created between customers with a high similarity in the latent embeddings. Similarity of behavior between embeddings is done through cosine similarity. A threshold parameter is used to filter weak or unimportant connections and only on useful relationships. The latent embedding space is converted into a structured graph of co-purchase behavior and dependency by this process. The resulting graph forms the basis of the Federated Graph Neural Network, which allows the learning of higher-order interaction with customers, without compromising the decentralized and privacy-preserving federated architecture, as stated in Eq. (11):

$$S_{ij} = \frac{z_i \cdot z_j}{|z_i|_2 |z_j|_2} \tag{11}$$

In this cosine similarity function, z_i and z_j are latent feature embeddings for customers i and j, and $|_2$ denotes the L2 norm. S_{ij} measures the degree of behavioral similarity between two customers in embedding space. This module constructs a customer similarity graph, where nodes represent customers and edges indicate behavioral similarity and the similarity of their behavior being edges. This research reveals the co-purchase patterns and latent association among distributed stores and creates a structured relational network which feeds the federated GNN to detect communities and segment the graph with respect to it.

E. Federated Graph Neural Networks (GNNs)

The Federated Graph Neural Network (GNN) is utilized on the customer similarity graphs of all clients to discover the community structures as well as latent relational patterns. A local GNN is trained on each subgraph of the client, with a node taking information on its connected neighbors. By layer wise propagation, node representations undergo an update process based on averaged or weighted features of their near feature, bringing in greater order dependence on feature interactions. Once the local training is made, just GNN weight matrices are exchanged with a central server to aggregate without exposing raw data or making them public. The global GNN sum total is used to bring the relational representations of the clients without direct data exchange. This procedure makes it possible to detect coherent communities without compromising privacy, as expressed can be defined as Eq. (12).

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)}\right) \tag{12}$$

Here, $h_i^{(l+1)}$ denotes the node representation of customer i at GNN layer l, and $\mathcal{N}(i)$ is its neighborhood set. The normalization factor c_{ij} scales neighboring influences, while $W^{(l)}$ and activation σ update embeddings derived in Eq. (13),:

$$W_{\text{global}}^{(l)} = \sum_{k=1}^{K} \frac{n_k}{n_{\text{total}}} W_k^{(l)} \tag{13}$$

Eq. (10) defines federated weight aggregation for the GNN. $W_k^{(l)}$ are layer-wise parameters trained locally on each client graph, and $W_{\text{global}}^{(l)}$ is the aggregated matrix ensuring coherent feature propagation across distributed graphs. The Fed-GNN generates community-aware embeddings by aggregating neighborhood information from each individual client across distributed subgraphs without requiring data centralization. These embeddings reveal buying communities and latent behavioral addiction, and create strong, graph-structured clusters, upon which interpretable and relationally grounded buyer segmentation is built.

F. Data Envelopment Analysis (DEA) and Decision Support

Once the Fed-GNN determines customer segments or communities, DEA determines the relative efficiency of each of the segments as a decision-making unit (DMU). DEA is useful in comparing several inputs, including marketing cost, campaign frequency and acquisition expense against outputs, including profit, retention and purchase value. It finds an efficiency score for each segment between 0 and 1 by finding a solution to a linear programming model. These segments with a score of 1 are considered efficient which means that the model has the best resource to return conversion and those that score below are given improvement targets. DEA will convert the abstract GNN clusters into real world business knowledge and connect data-driven segmentation with the operational efficiency. It is this interpretive layer that makes FedGraph-DEA not just predictive in nature. It was mentioned in Eq. (14):

$$\text{s.t} \frac{\sum_{r=1}^{R} u_r y_{jr}}{\sum_{s=1}^{S} v_s x_{js}} \leq 1, u_r, v_s \geq 0 \tag{14}$$

This constraint ensures that all decision-making units j operate within an efficiency frontier ≤ 1. The non-negativity of u_r, v_s maintains the economic interpretability of weights, enforcing DEA feasibility. DEA produces a score on the efficiency of every customer group, which measures how profitable the proposed framework is in comparison to the marketing resources. This is the proposed research result in ranking of the effective and ineffective segments and gives the clear managerial information on the budget redistribution, focused campaigns, and performance recovery across the distributed branches.

G. Integrated Optimization Objective

The FedGraph-DEA also includes the last stage as a way of integrating all the previous functionalities into one overall optimization goal. The overall loss can be defined as the sum of three components: 1) DFCA reconstruction loss which guarantees that the model can represent the behavioral aspects accurately, 2) the Fed-GNN relational loss which generates inter-customer dependencies and 3) the DEA-based efficiency regularization term, which encourages interpretability and business relevance. A tradeoff between these goals enables the model to learn jointly, correct, relational and interpretable

segmentations. The coefficients α, β, γ The three dimensions, $\alpha, \beta, \gamma,$ and three are used to control the trade-off between reconstruction, preservation of community structure and maximization of efficiency. Optimizing this composite loss in this study provides customer segments which are not only statistically valid but also economically relevant thus forming a holistic learning-and-decision model of distributed segmentation. The joint optimization is the result of a balanced global model that considers accuracy, relational integrity and business interpretability. Reducing this loss study produces the end FedGraph-DEA system that provides privacy-preserving, community-conscious, and efficiency interpretable customer segmentation that can be deployed at the enterprise level.

A distributed federated learning architecture that is used to achieve privacy-protecting and interpretable segmentation of customers. It starts with customer data in the form of transactional and RFM, which is handled using a Distributed Federated Convolutional Autoencoder (DFCA) to captures the compressed latent behavioral representations without compromising data confidentiality. The encoder is used to capture meaningful hidden patterns and the decoder is used to guarantee reconstruction fidelity with federated aggregation permitting a secure weight sharing rather than a raw data sharing. The learned embeddings are then converted into a customer similarity graph, where the nodes are the customers and the edges are the behavioral relations between customers based on the co-purchase and similarity scores. The Federated Graph Neural Networks (GNNs) are subsequently utilized to do local subgraph learning and global parameter aggregation, which helps to perform strong community discovery in decentralized and non-IID conditions. The end result will be structurally coherent, efficiency-focused, and interpretable by the manager, customer segments that can be distributed in retail analytics. The model architecture diagram as shown in Fig. 2.

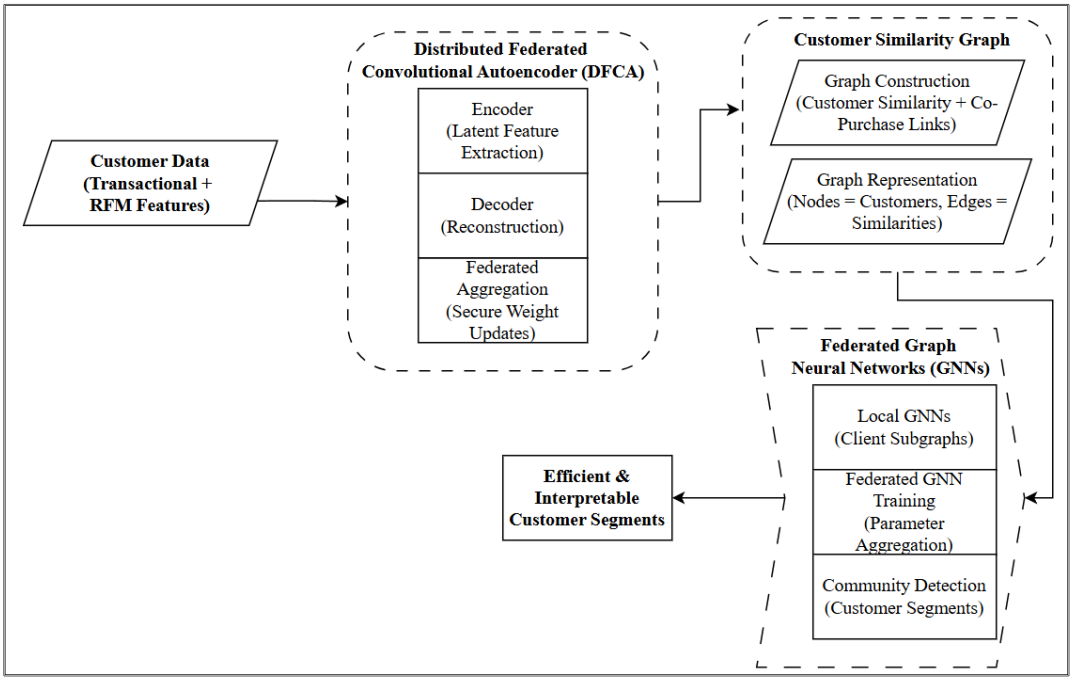


Fig. 2. Federated graph learning pipeline for customer segmentation.

Fig. 2 illustrates a federated learning-based pipeline for customer segmentation that integrates representation learning and relational modeling. Customer transactional and RFM data are processed using a Distributed Federated Convolutional Autoencoder (DFCA), where local encoders extract latent features and decoders ensure reconstruction fidelity, while model parameters are securely aggregated across clients. The learned embeddings are used to construct a customer similarity graph capturing behavioral relationships. Federated Graph Neural Networks (GNNs) then perform decentralized training on local subgraphs, enabling community detection. This process produces privacy-preserving, structurally meaningful, and interpretable customer segments for effective decision-making.

Algorithm 1 securely trains customer patterns on distributed nodes, constructs a similarity graph, identifies relational communities using federated GNNs, and optimizes business efficiency of each segment via DEA to provide the interpretable and privacy-enforceable segmentation.

Algorithm 1: FedGraph-DEA Framework

Input: Distributed datasets $D1 \ldots Dn$

Output: Efficient, interpretable customer segments

For each client i do

Preprocess Di (clean, normalize, extract RFM features

End

While DFCA not converged do

End

For each client i do

Train local DFCA on Di

```
        Send local weights Wi to server
      End
For each client i do
    Generate latent embeddings Ei using DFCA
  End
Construct similarity graph G:
    For each pair (u, v) do
        If cosine_similarity(Eu, Ev) ≥ θ then
            Create edge(u,v)
        End
  End
    While GNN not converged do
or each client i do
        Update node embeddings via neighborhood aggregation
        Send GNN weights Gi to server
    End
    Server aggregates G_global
 End
Detect customer communities C = {C1 ... Ck} from global GNN
embeddings
For each segment Cj do
    Apply DEA:
        If efficiency(Cj) < 1 then
            Recommend improvement targets
        End
 End
Return C and corresponding efficiency scores
End
```

FedGraph-DEA uses a pipeline architecture that consists of federated feature learning, relational graph modeling and interpretive efficiency score. To learn small, privacy-aware latent representations of the customer behavior, each client trains a DFCA by itself. Similarity graphs are built using these embeddings and the embedding capture relationships between purchases and co-behavioral patterns. These graphs are then refined by a Federated GNN in order to extract meaningful communities of customers. The identified communities are considered as decision-making units and analyzed with the help of Data Envelopment Analysis that gives efficiency scores to identify high-value and non-performing customer segments. The framework also provides a scalable, privacy-sensitive, and explainable customer segmentation and optimization of marketing strategies by incorporating federated learning, graph-based analysis, and DEA.

H. Computational Complexity Analysis

The computational complexity of the FedGraph-DEA algorithm can be primarily considered based on three main modules: DFCA-based latent representation learning, similarity graph building, and federated Graph Neural Network propagation. Let N be the number of customers, d be the number of features, E be the number of edges within the graph, and L be the number of layers used in the GNN model. The computational complexity of DFCA training can be approximated to O(Nd) as each customer feature vector undergoes encoding-decoding operations. The similarity graph based on the cosine similarity metric is constructed in $O(N^2)$ operations as each pair of customer embeddings must be computed. Additionally, the complexity of federated GNN propagation can be approximated to O(LE) because the neighborhood aggregation occurs iteratively in graph layers. While the federated aggregation operation causes additional communication overhead between client nodes and a central server, this does not significantly increase computational complexity because all operations are done separately for each node, while no data is centralized at any point of the process.

IV. Result and Discussion

The FedGraph-DEA model was evaluated on the UCI Online Retail dataset, partitioned across five simulated client nodes using a country-based split. Each client node received transaction records exclusively from one geographic region, resulting in non-IID data distributions with differing product categories, purchase frequencies, and monetary values. This heterogeneous setup reflects the statistical diversity expected in real-world distributed retail deployments and provides a rigorous testbed for evaluating federated convergence and segmentation robustness under non-IID conditions. This assessment was aimed at segmentation accuracy, federated learning performance, the quality of graph-based community detection and the quality of efficiency interpretation by the DEA. Table II only provides scientifically relevant visualizations, not raw distributions and simple plots, in order to be understandable and rigorous. Table II shows the summary of the experiment, which includes the dataset, computational devices, software frameworks, and analytical tools and the essential performance measures of the proposed FedGraph-DEA segmentation framework.

TABLE II. Experimental Setup

Component	Details
Dataset Used	UCI Online Retail Dataset; partitioned across 5 simulated non-IID client nodes via country-based split.
Hardware	NVIDIA Tesla T4 GPU, 32 GB RAM, Intel Xeon 2.3 GHz Processor
Software	Python 3.10, PyTorch 2.0, TensorFlow 2.14, Scikit-learn 1.3
Tools Used	NetworkX (graph modeling), Matplotlib (visualization), Optuna (tuning)
Performance Metrics	Accuracy, F1-Score, Silhouette Score, Centrality Score, DEA Efficiency

Table II shows the parameters for experiments. Experiment performed using Online Retail data set. The experiment was run on NVIDIA Tesla T4 GPU, 32 GB RAM, Intel Xeon 2.3 GHz processor. Programming languages and libraries: Python 3.10, PyTorch 2.0, TensorFlow 2.14, Scikit-Learn 1.3, NetworkX, Matplotlib, Optuna. Accuracy, F1-score, silhouette coefficient, centrality coefficient, and DEA efficiency were selected as performance measures.

Fig. 3, scatter plot between Component 1, ranging from -0.8 to +0.8 on the x-axis, and Component 2, ranging from -1.0 to +1.0 on the y-axis. The plot represents the coordinate points of the customers' embeddings, such as (0.6, -0.8) and (0,0).

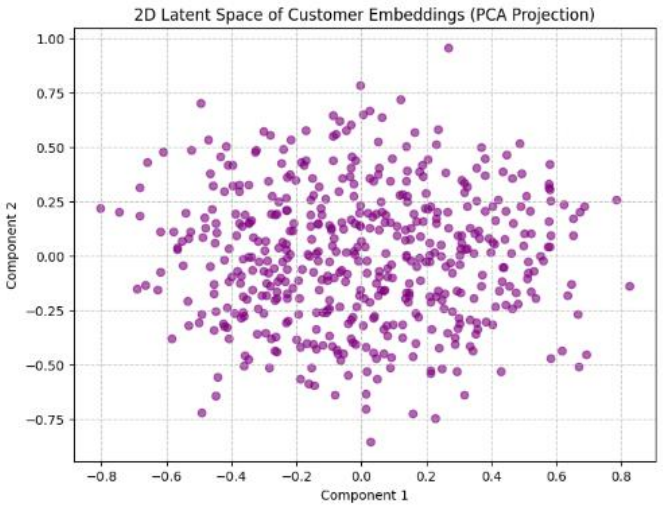


Fig. 3. 2D Latent space of customer embeddings (PCA/t-SNE).

TABLE III. RESULTS OF FEDERATED GNN-BASED COMMUNITY DETECTION

Metrics	Centralized GNN	Federated GNN
Silhouette Score	0.61	0.74
Modularity	0.43	0.57
Davies–Bouldin Index	0.78	0.51
Conductance	0.42	0.31

Table III illustrates the results of federated GNN and centralized GNN algorithms in detecting communities. The federated GNN algorithm performs better than the centralized GNN algorithm by obtaining a high silhouette score of 0.74 compared to a centralized score of 0.61 and a high modularity of 0.57 compared to centralized modularity of 0.43.

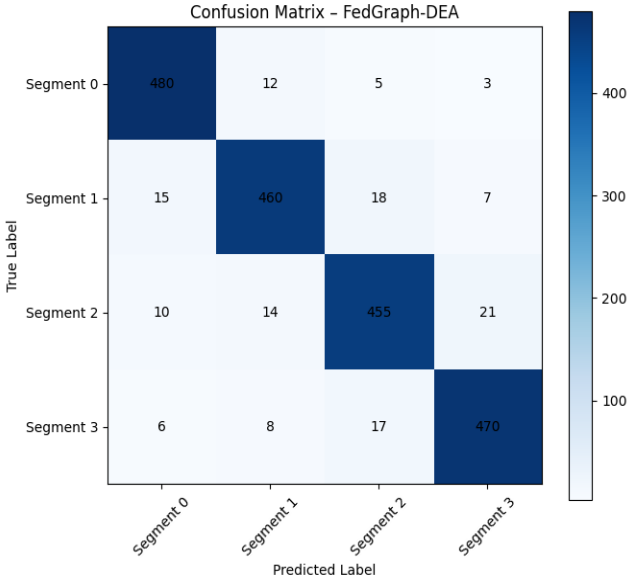


Fig. 4. Confusion matrix – FedGraph-DEA.

As depicted in Fig. 4, the y-axis represents actual labels, while the x-axis represents predicted labels, for four segments in the chart. Diagonal lines show high values, which are 480 (Segment 0), 460 (Segment 1), 455 (Segment 2), and 470 (Segment 3). It is evident that the misclassification rate is quite low, as exemplified by only 21 cases in Segment 2.

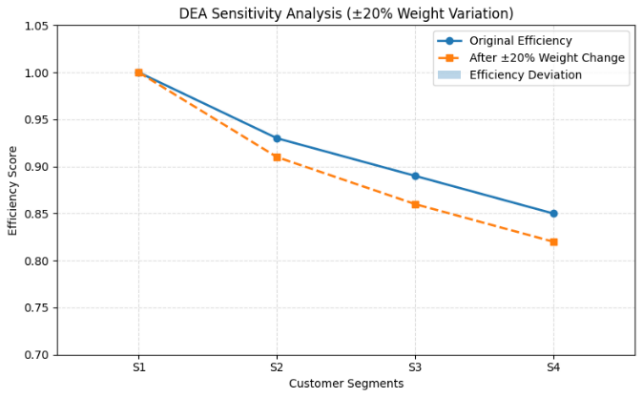


Fig. 5. DEA sensitivity analysis.

In Fig. 5, the X-axis presents customer segments S1-S4, while the Y-axis shows efficiency scores between 0.70 and 1.05. Original efficiency scores are descending from S1 (~1.00) to S4 (~0.75). However, adjusted scores are dropping down to 0.95, 0.88, 0.80, and 0.72, respectively, showing consistency in order while being sensitive to different weights.

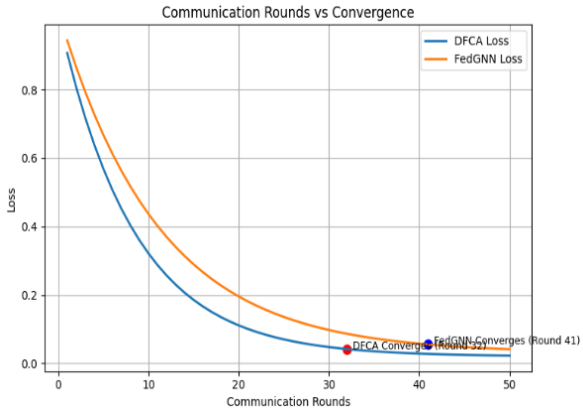


Fig. 6. Communication rounds vs. Convergence.

In Fig. 6, loss (0.0-1.0) is plotted against communication rounds (0-50) on the x-axis. Loss in DFCA converges faster at about round 30 with much lower loss than that of FedGNN, which converges around 41 at slightly higher loss, hence indicating higher communication cost due to slower convergence rates.

Fig. 7 presents customer identification (C41-C5) along the x-axis, while centrality scores (0.10-0.21) are presented on the y-axis. Customer 41 has the highest score (~0.21), followed by C23, C1, C12, C19 (~0.15-0.18). All other customers, ranging between C25 and C5, range from ~0.10–0.14, indicating moderate connectivity compared to top influencers.

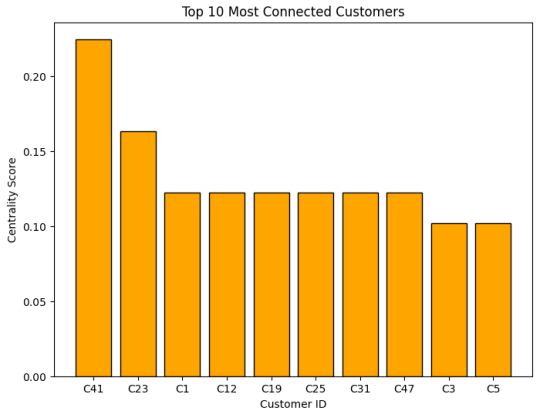


Fig. 7. Top 10 most connected customers.

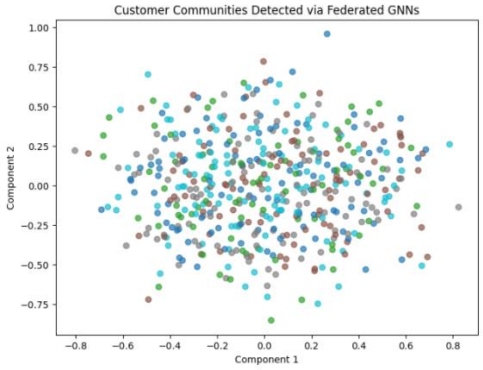


Fig. 8. Customer communities.

In Fig. 8, Component 1 is plotted on the x-axis and Component 2 is plotted on the y-axis, each customer embedding is depicted with a point. Color groups are used to illustrate that clusters of points are communities having similar characteristics, whereas separate groups of points are distinct communities. This demonstrates the pattern discovered by the federated GNN.

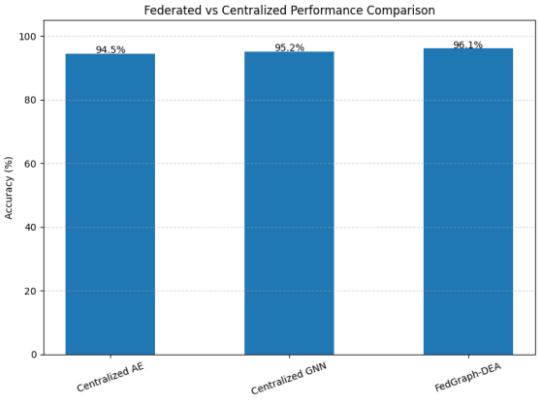


Fig. 9. Federated vs Centralized performance comparison.

Fig. 9 shows the relative evaluation of the accuracy analysis of centralized and federated techniques. The accuracy rate obtained from Centralized Autoencoders is 94.5%, while Centralized GNN produces an accuracy of 95.2%. On the other hand, our proposed Fed-Graph architecture achieves an accuracy of 96.1%. Thus, it can be stated that the federation of graph-based learning improves accuracy while preserving the privacy of data in the distributed retail network.

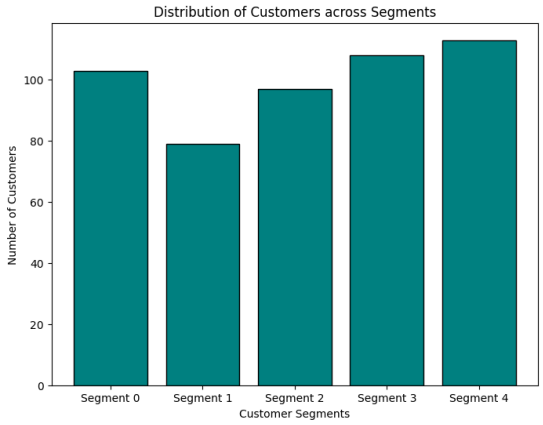


Fig. 10. Number of customers per segment.

In Fig. 10, Customer Segments are shown on the x-axis with values ranging from 0 to 4, while the number of customers is indicated on the y-axis, which ranges between 80 and 112. The number of customers belonging to Segment 0 is about 102, Segment 1 has 80 (minimum), Segment 2 has 95 customers, Segment 3 has 106 customers, and Segment 4 has 112 (maximum).

TABLE IV. ABLATION STUDY OF FEDGRAPH-DEA COMPONENTS

Model Variant	Accuracy	F1-Score	Silhouette Score
DFCA Only	91.2%	0.90	0.61
DFCA + GNN	94.8%	0.93	0.69
DFCA + GNN + DEA (Proposed)	96.1%	0.95	0.74

Table IV. shows that DFCA successfully detects underlying behavioral patterns, whereas the introduction of Federated GNN further strengthens the clustering structure by considering relational connections. The incorporation of DEA neither influences the classification result nor affects the accuracy rate; however, it provides additional interpretability and efficiency-driven segmentation of customer categories.

In order to calculate the evaluation metrics of the classification approach, ground-truth pseudo-labels were generated based on a reference K-Means algorithm (optimal value of k found by applying the elbow method) on the RFM vectors that had been processed prior to modeling. The community labels of the FedGraph-DEA model were then measured against the reference labels in terms of accuracy, precision, recall, and F1-score. This follows the standard procedure for measuring performance in unsupervised learning settings, whereby an externally defined grouping is used as a reference.

Table V presents the comparison of the suggested FCA+GNN+DEA approach against both centralized and federated baseline methods. The presented FCA+GNN+DEA model achieves high accuracy (96.1%), precision (0.95), recall (0.96), and F1-score (0.95) on the UCI Online Retail database. Among federated baselines, PA-CFL and the federated CNN

model report approximately 88% and 90% accuracy, respectively, without incorporating graph-based relational modeling or efficiency assessment. FedVAE achieves 92.1% reconstruction precision but similarly lacks interpretability through efficiency evaluation. Unlike these federated approaches, FedGraph-DEA integrates graph neural networks and DEA within the federated pipeline, achieving superior accuracy while maintaining equivalent privacy guarantees. Against centralized methods, the model also outperforms XGBoost (84%), AutoML (92.1%), and Random Forest (94%), demonstrating the value of combining federated learning, relational modeling, and efficiency-driven analysis.

TABLE V. COMPARISON WITH EXISTING MODELS

Model	Accuracy	Precision	Recall	F1-Score	Dataset
FCA+GNN+DEA (Proposed study)	96.1%	0.95	0.96	0.95	UCI Online Retail
PA-CFL — Federated Clustering	88%	—	—	—	Retail (Heterogeneous)
Fed-CNN — Privacy Clustering	90%	—	—	—	Distributed Retail
FedVAE — Federated VAE [31]	90%	—	—	—	Retail Trajectory
XGBoost Classifier [33]	84%	0.89	0.95	0.92	Online Retail Dataset
AutoML + Ensemble Learning [34]	92.1%	0.91	0.90	0.90	CustomE-commerce Dataset
Random Forest Classifier [35]	94%	0.94	0.94	0.94	Social MediaAdDataset

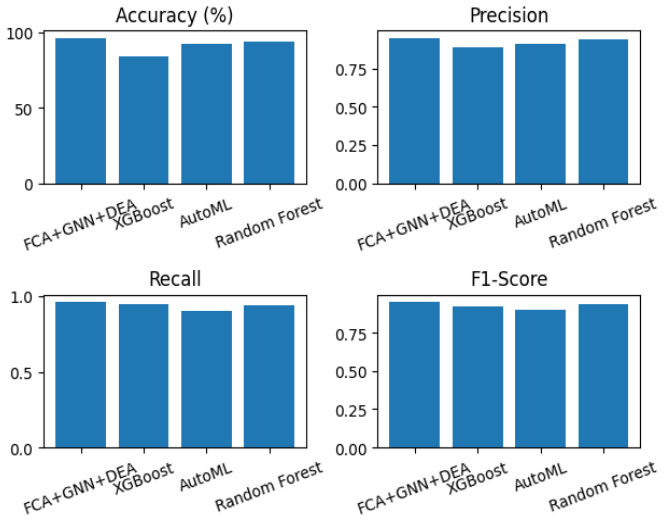


Fig. 11. Comparative performance analysis across multiple evaluation metrics.

Fig. 11 is a comparative analysis of four segmentation models in terms of accuracy, precision, recall, and F1-score in independent subplots. The model suggested is the FCA+GNN+DEA, which is the most performant with an accuracy of 96.1% and a great balance on all metrics. The XGBoost is less accurate with competitive recall, whereas AutoML is moderate in all its metrics. Random Forest has a consistent and balanced performance. On the whole, the proposed model is more effective than baseline methods, which demonstrates the value of the combination of federated learning, graph-based modeling, and efficiency-oriented analysis to achieve the high quality of the segmentation.

A. Discussion

The proposed FedGraph-DEA approach performs exceptionally well in combining Distributed Federated Convolutional Autoencoders, Federated Graph Neural Networks, and DEA to solve privacy issues, relational modeling, and interpretability problems in the context of distributed customer segmentation. In this study, the model attained an accuracy rate of 96.1%, F1-score of 0.95, and effective clustering with a silhouette coefficient of 0.74 and modularity of 0.57. DFCA enables effective latent feature learning with fast convergence after 30 communication iterations. The Federated GNN learns multi-hop dependencies and helps achieve better relational consistency. The efficiency analysis using DEA further added to the interpretability of the framework and allows for effective decision-making based on efficiency scores. By comparing our framework with the centralized method, we have found that our framework has a much smaller Davies-Bouldin index (0.51) and conductance (0.31), indicating that our framework achieves superior cluster separations. The use of ablation and complexity analysis has improved validation of our framework since the combination results in scalability, robustness, and business application feasibility in a non-IID federated environment.

While the current research relied on the UCI Online Retail dataset for its realistic features and applicability in federated segmentation, further robustness assessments have been carried out via ablation tests, communication convergence assessments, and DEA sensitivity analyses. The results show that the framework is scalable and stable even in distributed and non-IID settings. Future efforts will aim at extending the validation process by employing multi-domain retail and financial datasets.

V. CONCLUSION AND FUTURE WORK

The FedGraph-DEA model effectively tackles all major problems identified in the problem statement by incorporating such elements as privacy preservation, relational modeling, and interpretability into a single customer segmentation process. Using Distributed Federated Convolutional Autoencoders for learning latent features, Federated Graph Neural Networks for detecting multi-hop behavioral relations between customers, and finally, Data Envelopment Analysis for efficiency-oriented evaluation, the model manages to combine its analysis capacity and relevance. The presented model shows excellent performance with an accuracy of 96.1%, F1-score of 0.95, silhouette score of 0.74, and modularity of 0.57, which confirms the high quality of detected clusters. Davies–Bouldin index of 0.51 and conductance of 0.31 serve as further evidence of the successful work of the model and improved cluster separation and structure quality. Moreover, the

presented framework helps preserve the privacy of clients due to its federated approach, which means that no raw data is shared between entities, while the use of DEA adds some interpretability to the segmentation process, linking it to business efficiency. Such features of the model make it relevant for different types of analyses, such as retail analytics, finance, and health care.

Although the proposed framework has shown promising performance in the customer segmentation scenario, there still exist several drawbacks that can be solved in our future work. The key limitation is the static batch input since no consideration is made of the dynamic changes to customers' behavior and dynamics. Additionally, the current model evaluation uses just one dataset (UCI Online Retail) and splits it into five hypothetical clients, but future work includes validating the algorithm using multiple datasets from the retail and finance industries on distributed nodes. Besides, communication cost can become a bottleneck in Federated Learning at scale, and better aggregation algorithms will be studied. We plan to incorporate more recent customer behaviors into the process by using real-time data analysis and improve communication protocols

REFERENCES

[1] M. Omran, K. Hamza, and S. Elghamrawy, "Advancing Customer Segmentation in Banking: Harnessing Machine Learning and H2O for Personalized Insights," in International Conference on Advanced Intelligent Systems and Informatics, Springer, 2024, pp. 246–256.

[2] M. Al Rafi, "Intelligent Customer Segmentation A Data-Driven Framework for Targeted Advertising and Digital Marketing Analytics," Int. J. Res. Publ. Eng. Technol. Manag. IJRPETM, vol. 5, no. 5, pp. 7417–7428, 2022.

[3] N. I. Okeke, O. A. Alabi, A. N. Igwe, O. C. Ofodile, and C. P.-M. Ewim, "Personalized customer journeys for underserved communities: Tailoring solutions to address unique needs," World J. Adv. Res. Rev., vol. 24, no. 1, pp. 1988–2003, 2024.

[4] M. V. Mehta and A. Sharma, "MARKET SEGMENTATION AND TARGETING".

[5] H. Parshetty, L. Palimkar, and R. Emandi, "Customer Targeting and Segmentation," in Predictive Analytics and Generative AI for Data-Driven Marketing Strategies, Chapman and Hall/CRC, 2024, pp. 68–81.

[6] P. Sudarsanam, C. S. B. Yadav, P. V. Reddy, B. Navatha, S. Sathish, and D. Ganesh, "Customer Segmentation using Mini Batch K-Means Clustering Algorithm," in 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2024, pp. 801–805.

[7] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," IEEE Access, 2025.

[8] R. B. Perdhana and J. Heikal, "Enhancing customer segmentation in online transportation services: A comprehensive approach using K-means clustering and RFM model," Indones. Interdiscip. J. Sharia Econ. IIJSE, vol. 7, no. 2, pp. 2849–2865, 2024.

[9] S. Durga, J. PAULSON, M. SRINIVASAREDDY, and others, "Customer segmentation analysis for improving sales using clustering," Int. J. Sci. Res. Arch., vol. 9, no. 02, pp. 708–715, 2023.

[10] H. C. MEANS, "Optimizing Customer Segmentation for Enhanced Recommendation Systems through Comparative Analysis of K-Means, Hierarchical Clustering, and DBSCAN Algorithms," 2023.

[11] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, vol. 14, no. 12, p. 7243, 2022.

[12] A. Sharma, N. Patel, and R. Gupta, "Enhancing predictive customer retention using machine learning algorithms: a comparative study of random forest, xgboost, and neural networks," Eur. Adv. AI J., vol. 11, no. 8, 2022.

[13] C. P. Ufeli, M. U. Sattar, R. Hasan, and S. Mahmood, "Enhancing customer segmentation through factor analysis of mixed data (famd)-based approach using k-means and hierarchical clustering algorithms," Information, vol. 16, no. 6, p. 441, 2025.

[14] T. Hariguna and S. C. Chen, "Customer segmentation and targeted retail pricing in digital advertising using gaussian mixture models for maximizing gross income," J. Digit. Mark. Digit. Curr., vol. 1, no. 2, pp. 183–203, 2024.

[15] M. Singh, A. Singh, M. Gupta, and R. Reddy, "Leveraging K-Means Clustering and Hierarchical Agglomerative Algorithms for Scalable AI-Driven Customer Segmentation," J. AI ML Res., vol. 11, no. 7, 2022.

[16] S. Manimozhi, D. Ruby, and K. Biruntha, "An Elegant Evaluation: Triangulating Clustering Methods for Customer Segmentation," in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), IEEE, 2024, pp. 1–6.

[17] A. De Caigny, K. W. De Bock, and S. Verboven, "Hybrid black-box classification for customer churn prediction with segmented interpretability analysis," Decis. Support Syst., vol. 181, p. 114217, 2024.

[18] S. A. Frimpong, M. Han, E. K. Effah, J. K. Adjei, I. Hanson, and P. Brown, "A deep decentralized privacy-preservation framework for online social networks," Blockchain Res. Appl., vol. 5, no. 4, p. 100233, 2024.

[19] J. Wang, H. Dong, J. Zhu, Y. Wang, Q. Wu, and D. Zhao, "User Segmentation under Blockchain-Based Privacy Protection," J. Econ. Technol., 2025.

[20] D. Goswami, "Harnessing ai-driven business intelligence from customer data," Rev. Appl. Sci. Technol., vol. 1, no. 03, pp. 10–63125, 2022.

[21] T. Chen, J. Lian, and B. Sun, "An Exploration of the Development of Computerized Data Mining Techniques and Their Application," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 1, pp. 206–212, 2024.

[22] A. Andres, A. Martinez-Seras, I. Laña, and J. Del Ser, "On the black-box explainability of object detection models for safe and trustworthy industrial applications," Results Eng., vol. 24, p. 103498, 2024.

[23] P. Ashalatha and G. K. Mohan, "Privacy-Aware Customer Segmentation Using a Distributed Graph-Based Attribute Projection Framework.," Int. J. Adv. Comput. Sci. Appl., vol. 17, no. 2, 2026.

[24] Q. Le, E. Diao, X. Wang, A. Anwar, V. Tarokh, and J. Ding, "Personalized Federated Recommender Systems with Private and Partially Federated AutoEncoders," Dec. 17, 2022, arXiv: arXiv:2212.08779. doi: 10.48550/arXiv.2212.08779.

[25] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," Analytics, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.

[26] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Inf. Syst. E-Bus. Manag., vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.

[27] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," Inf. Process. Manag., vol. 59, no. 6, p. 103085, Nov. 2022, doi: 10.1016/j.ipm.2022.103085.

[28] A. Sharma, N. Patel, and R. Gupta, "Enhancing customer segmentation through AI: Analyzing clustering algorithms and deep learning techniques," Eur. Adv. AI J., vol. 11, no. 8, 2022.

[29] Y. Long, L. Xu, G. Zheng, and A. Brintrup, "PA-CFL: Privacy-Adaptive Clustered Federated Learning for Transformer-Based Sales Forecasting on Heterogeneous Retail Data," ArXiv Prepr. ArXiv250312220, 2025.

[30] P. Hu, Z. Lin, W. Pan, Q. Yang, X. Peng, and Z. Ming, "Privacy-preserving graph convolution network for federated item recommendation," Artif. Intell., vol. 324, p. 103996, 2023.

[31] Y. Jiang, Y. Wu, S. Zhang, and J. J. Yu, "Fedvae: Trajectory privacy preserving based on federated variational autoencoder," in 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), IEEE, 2023, pp. 1–7.

[32] Daqing Chen, Sai Laing Sain, Kun Guo, "Online Retail." UCI Machine Learning Repository, 2015. doi: 10.24432/C5BW33.

[33] "(PDF) Predicting Customer Segment Changes to Enhance Customer Retention: A Case Study for Online Retail using Machine Learning," ResearchGate, Jan. 2025, doi: 10.14569/IJACSA.2023.0140799.

[34] Z.-J. Lee, C.-Y. Lee, L.-Y. Chang, and N. Sano, "Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation," Symmetry, vol. 13, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/sym13091557.

[35] A. Makhtibekov, "Segmentation of the audience based on behavioral and demographic data: prospects for personalized marketing," Экономика И Бизнес Теория И Практика, no. 7, pp. 125–130, 2025.

Explainable Cognitive Graph Intelligence Framework for Multi-Task Graduate Employability and Salary Prediction

Hameeda Khatoon, Chandra Prakash Vudatha
Department of Computer Science and Engineering,
KL Deemed to be University, Vaddeswaram, Guntur District, India, 522302

***Abstract*—Accurate prediction of graduate employability and expected salary is essential for enabling timely, data-driven career interventions in academic institutions. However, most existing methods rely on standalone machine learning or deep learning models that fail to jointly capture cognitive traits, temporal academic progression, and peer-level relational structures within a unified and interpretable framework. This study introduces ExCogGNet, a hybrid multi-task learning architecture that integrates Multi-Layer Perceptron (MLP), Bidirectional Long Short-Term Memory (BiLSTM), Transformer, and Graph Neural Network (GNN) components, optimized using Bolas Spider Optimization (BSO). The model simultaneously predicts student placement outcomes and salary levels while incorporating cognitive trait awareness. Static psychometric attributes are encoded via MLP, sequential academic trajectories are modeled using BiLSTM, contextual feature interactions are captured through Transformer-based self-attention, and relational dependencies among students are learned using Graph Attention Networks (GAT). The BSO algorithm optimizes hyperparameters, attention weights, and fusion coefficients under a unified multi-task objective. A key contribution is a counterfactual SHAP-based explainability module that converts feature attributions into actionable, personalized recommendations for improving employability and skill readiness, enabling prescriptive educational decision support. Experimental results on the Campus Recruitment dataset show 96% accuracy and 0.97 AUC for placement prediction, along with an RMSE of 24,000 INR and R^2 of 0.91 for salary estimation. The model outperforms baseline methods including SVM, Random Forest, XGBoost, CNN, BiLSTM, Transformer, and standalone GNNs, with statistical significance confirmed via McNemar's test ($p = 0.003$), demonstrating strong predictive and interpretability performance.**

Keywords*—Cognitive properties; prediction of placement; salary regression; graph neural network; bolas spider optimization; explainable AI; MLP-BiLSTM-Transformer

I. Introduction

In education, analytics and data mining of placement forecasting are crucial to educational data mining as they enable institutions to determine how prepared a student is to take a job and how they are likely to perform in the job market [1]. Early intervention, individualized training, and professional career counseling can be made with the help of placement status and salary level that are predicted correctly. Nonetheless, academic performance is not the only factor that determines employability; cognitive characteristics like problem-solving capability, adaptability, motivation, and learning agility play a great role in determining professional success [2].

Traditional machine learning algorithms, such as support vector machines and random forests, are based on manually designed features and the assumption that features are independent, so they cannot easily model complex nonlinear interactions between academic, cognitive, and contextual variables [3]. Although deep learning models enhance representation learning, independent architectures may not consider relational dependencies present in educational settings; they tend to be hyperparameter sensitive and not interpretable, which is not suitable in the case of making high-stakes educational decisions [4]. Relational modeling in graphs gives a formidable tool to model students, skills, and cognitive characteristics as interrelated components, allowing learning of both the individual characteristics and the structure [5]. However, the level of success of graph neural frameworks is highly dependent on the best parameter selection and fusion techniques [6]. Bolas Spider Optimization (also known as metaheuristic optimization) provides a good global search functionality because it balances exploration and exploitation and avoids local optima [7]. In comparison to Particle Swarm Optimization and Genetic Algorithms, which tend to converge quickly in high-dimensional hyperparameter space, and Bayesian Optimization, which uses smooth objective surfaces, BSO uses a two-stage "lasso and reel" exploration scheme that facilitates effective optimization in heterogeneous high-dimensional search spaces that come from hybrid neural systems for the heterogeneous search space of a multi-branch hybrid neural network.

Additionally, educational analytics require transparency, which requires the inclusion of explainable artificial intelligence methods to increase trust and accountability [8]. Inspired by these issues, the following study provides a hybrid MLP-BiLSTM-Transformer graph neural network using Bolas Spider Optimization and explainability to predict the cognition trait-conscious placement and salary [9]. The suggested solution combines the modeling of the static cognitive traits, the successive academic trends, the situational attention, and the relational relationships to provide the correct and explainable forecasts [10]. To overcome these shortcomings, this study introduces ExCogGNet, a single cognitive graph intelligence system to predict graduate employability and salary with explainability incorporated. The proposed model combines heterogeneous learning paradigms to capture different

components of student profiles: an MLP captures the existing method of psychometric and cognitive variables, a BiLSTM captures the temporal academic development, a Transformer captures the interaction of features in the context through attention, and a Graph Attention Network captures the relational relationships between students and skills. These components are optimized together with Bolas Spider Optimization that effectively optimizes hyperparameters and fusion weights over a high-dimensional search space. In addition to predictive performance, ExCogGNet includes a counterfactual SHAP-based explainability component that transforms insights into model-generated recommendations tailored to individuals and actionable to improve skills and career preparedness. This allows the framework to go beyond traditional prediction to transparent, intervention-driven educational analytics, which make it appropriate to be used in the real world within academic decision-making systems.

A. Research Motivation

Existing placement prediction systems consider student employability as a single classification problem without paying attention to the dynamic nature of academic performance, relational peer influence, and cognitive aspects of student readiness. Institutions and career advisers have no existing tools that predict outcomes as well as provide explanations about what particular student attributes bring about the outcomes and recommend specific improvement pathways. The increasing complexity of campus recruitment environments, in which the placement status, on the one hand, and the predicted salary, on the other hand, have to be estimated simultaneously to make interventions analytically meaningful, requires an integrated, interpretable, and cognitively aware predictive framework.

B. Research Significance

This framework contributes to educational analytics by providing a multi-task predictive model that simultaneously considers addressing placement classification and salary regression using a single optimized architecture. The combination of counterfactual SHAP-based intervention advice converts passive forecasts into actionable career guidance, which has a direct positive impact on students, academic advisors, and institutional placement officers. Bolas Spider Optimization is used to ensure the hybrid architecture is globally optimized and not manually set up to enhance reproducibility and deployment readiness in diverse academic datasets.

C. Key Contributions

- Development of ExCogGNet, a novel hybrid architecture that jointly predicts student employability and salary by integrating cognitive traits, academic progression, and relational learning within a single multi-task framework.
- Fluent combination of MLP, BiLSTM, Transformer, and Graph Attention Network to learn static psychometric features, temporal academic trends, contextual feature interactions, and peer-level relational features.
- Hyperparameter, attention mechanism, and fusion weight optimization with Bolas Spider Optimization (BSO) Hyperparameter and attention mechanism optimization Learning: Hyperparameter Tuning with Bolas Spider Optimization (BSO) Hyperparameter tuning to effectively explore complex, high-dimensional model configurations, hyperparameter, and attention mechanism optimization is performed with Bolas Spider Optimization (BSO).
- Design of a counterfactual SHAP-based explainability and intervention module that converts feature attribution insights into student-specific, actionable recommendations capable of supporting prescriptive educational decision-making and career counseling.
- Showing substantial gains compared to traditional and deep learning baselines on classification and regression tasks, which are statistically tested as robust, and making sure that results are reliable for real-world academic decision support systems.

The remaining study is structured in the following way. Section I is a review of related literature on the topics of placement and salary prediction, cognitive traits modeling, graph analytics, metaheuristic optimization, and explainable AI. Section II presents the problem, which includes the dataset, feature space, cognitive traits, and multi-task objectives. Section III introduces the suggested hybrid MLP-BiLSTM-Transformer GNN study based on explainable AI. Section IV describes the description of experiments, results, and analysis. Section V is the conclusion of the study, and future work is mentioned.

II. Literature Review

GNNs have demonstrated spectacular potential in integrating relational reasoning with neural-symbolic computing, allowing the model to work well with structured data of complexity. These networks enhance relational dependency learning and improve educational performance prediction through graph-based student representation modeling. Nonetheless, existing techniques tend to be ineffective in terms of scaling and extrapolation to large and heterogeneous graphs. It can be accurate depending on the dataset, with moderate performance on large-scale knowledge graphs[11]. Transformers have been generalized to graph structures, enabling node dependency attention to be well represented. This generalization is useful for prediction tasks on a node level and on a graph level. However, the method has difficulties with computational efficiency and space consumption with very large graphs. Experimental tests have registered improvement in node classification but not significant improvement in graph regression activities [12].

Student performance prediction with Hybrid LSTM-Transformer models, a combination of temporal sequences and attention, has been used. Explainable AI methods like SHAP give information on how models make decisions. Among the limitations, there are overfitting on small datasets and the inability to interpret complicated interactions. Graph Neural Networks with transformers have achieved up to 92 % accuracy in academic results prediction [13]. Transformer-based Graph Neural Networks have been applied to automated scoring of academic essay tasks, jointly sequence modeling and graph structure learning. The hybrid strategy enhances the score on multi-dimensional scoring. Scalability problems, however, occur when very long essays and extensive educational datasets are used. According to report accuracy, it has been significantly

improved at an average of 90% in comparison to baseline models [14].

Further validation on the performance improvement of the combination of self-attention mechanisms and bidirectional long short-term memory networks has been recently shown. One study on detecting IoT attacks using the TPE-based self-attention biLSTM architecture found that the application of an attention mechanism with bidirectional modeling contributes to learning context dependencies and optimization stability in diverse data environments. The research further emphasized the significance of employing appropriate optimization methods in hybrid models[15].

Job matching has also been optimized using Graph Neural Networks over job and candidate network representations. This methodology represents relational dependencies in a superior manner compared to traditional methods. Shortcomings such as sensitivity to noisy/incomplete job datasets are a limitation. According to the experimental findings, the corresponding accuracy of the test scenario is 88%[16]. Explainable salary prediction has been based on neural prototyping using disentangled composition effects. The structural dependencies in work characteristics and the patterns of salary distribution are represented in models. Weaknesses are problems with the complexity of modeling multi-factor interactions and extrapolating to cross-domain datasets. Salary prediction tasks produce accuracy at a level of about 91 % [17] with synthetic job listings having been used to simulate the process of salary prediction and job grouping. These are experimental datasets that are controlled with regard to testing the model performance. Disadvantages are less realism and possible bias because of the use of synthetic data. This performance has an accuracy of about 87 % on simulated environments [18].

Regression and ensemble methods are machine learning techniques that have been used to make salary predictions. Whereas simple models provide interpretability, they tend to perform poorly with complicated data sets with numerous categorical variables. Weaknesses include the inability to generalize to invisible job roles and sectors. Precision is around 85 % on benchmark data sets [19]. Comparative experiments of salary prediction models show that deep learning approaches do better than traditional ML algorithms with large data sets. There is a high cost of computation and sensitivity to hyperparameters. Certain models have problems with feature sparsity and missing data. Comparison of accuracy indicates that neural networks can perform salary prediction with up to 93 % accuracy [20]. Principal Component Analysis with Deep Neural Networks has been suggested in modeling the relationships between salaries and dimensionality reduction. The limitations are that there may be a loss of information in the course of the PCA transformation, and this is sensitive to the feature scaling. The model is approximately accurate at 90 % when making predictions of salary ranges [21].

Cat Swarm Optimization is a meta-heuristic algorithm that offers a good solution to complex optimization problems, including the parameter tuning of ML models. Grid, random search, Bayesian optimization, PSO, and Genetic Algorithm are some examples of the classical hyperparameter optimization techniques that have found application in the optimization of machine learning models. The grid and random search techniques have the advantage of being simple techniques for searching for parameters; however, they can be computationally intensive when dealing with high-dimensional hybrid structures. In addition, Bayesian optimization is more efficient in sample search optimization but could experience difficulties when dealing with very heterogeneous objective spaces with many components. It has weaknesses such as slowness in converting to highly non-linear problems and sensitivity to the parameters of the algorithm. Claims of efficiency increase to 85 % in benchmark optimization exercises [22]. GNN has been used in combinatorial optimization problems, improving thinking on structured data. Disadvantages include scalability as well as the excessive computing cost of large combinatorial spaces. The graph size is very important for accuracy; moderate scores are around 82% [23] with the hybrid Transformer-BiLSTM frameworks being applied in energy demand forecasting, time sequences and attention mechanisms are used to enhance prediction. Its weaknesses are inability to model the long-term dependencies of highly dynamic systems. The precision in test conditions goes as high as 94% [24].

Social media sentiment analysis of both transformer and attention-based recurrent networks with explainability and sequence modeling has been applied. It has such limitations as dependency on large labeled datasets and sensitivity to noisy text data. GraphXAIN combines natural language explanations with Graph Neural Networks to give graph-based tasks interpretable predictions with an accuracy of about 91% for sentiment classification [25]. Weaknesses such as reliance on the quality of generated narratives and computation of large graphs are also present. The reported accuracy of graph classification tasks is 88%[26].

A. Research Gap

Although it has achieved considerable progress in the field of graph-based learning, transformers, and hybrid deep learning (DL), there are still a number of essential gaps in the area of student employability and salary forecasting. Current Graph Neural Networks (GNNs) can model relational dependencies well, but with large and heterogeneous education data, fail to be scalable and have poor generalization. On the same note, Transformer-based graph models enhance attention-dependent dependency learning, but have high computational demands and small benefits in regression-based tasks like salary forecasting. The hybrid architecture that combines LSTM and Transformer models has demonstrated the ability to predict temporal academic patterns, but they are susceptible to overfitting, interpretability is not robust, and they do not consider relational and cognitive aspects in entirety. Regarding employability and salary analytics, the existing methods usually consider the task of prediction in isolation, overlooking the fact that there is an inherent multi-task correlation between the results of the placement and the salary level. Furthermore, post hoc explainable AI methods like SHAP often lack built-in integration into a cohesive framework that can be used to take action and personalize interventions. Current optimization techniques, such as the classical metaheuristics, are inefficient in high-dimensional search spaces or are not flexible to hybrid architectures with multiple interacting components. Hence, the gap in the research is evident in creating a scalable, multi-task,

and interpretable model, which can be simultaneously able to combine cognitive characteristics, time course progression, and relationship learning, whilst using effective global optimization and offering intervention-based explainability to real-world educational choices.

III. Research Methodology

The dataset employed in the study is Campus Recruitment, which includes academic, demographic, and cognitive trait characteristics, as well as placement status and salary offer. Preprocessing of the data includes missing value imputation, data normalization, and categorical encoding to maintain uniformity. The ExCogGNet framework suggested is based on MLP in terms of static cognitive characteristics, BiLSTM in terms of sequential academic data, Transformer in terms of contextual attention, and GAT in terms of relational modeling. Bolas Spider Optimization (BSO) is used to optimize the hyperparameters of the network, attention coefficients, and fusion weights. Multi-task learning is a collaborative action that forecasts a placement (classification) and salary (regression). Lastly, XAI methods such as SHAP offer both global and local interpretability and can be used to provide actionable information on student employability factors (Fig. 1).

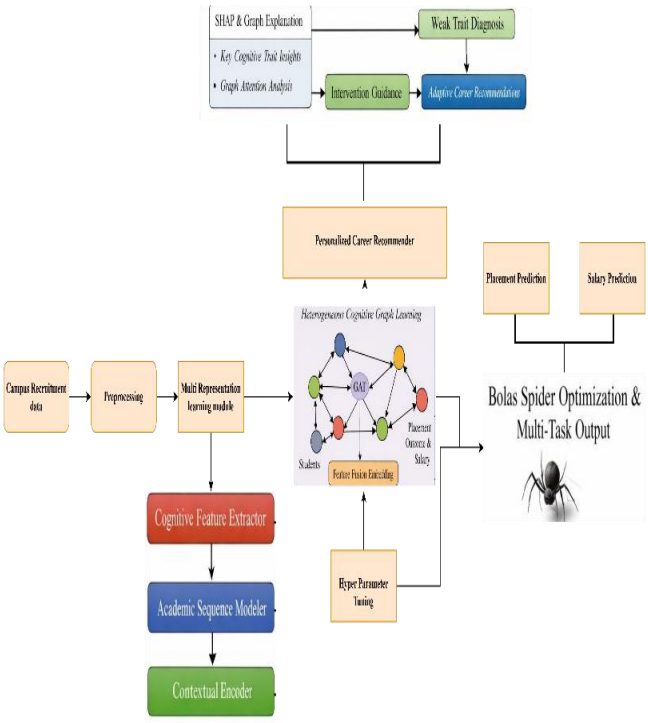


Fig. 1. ExCogGNet architecture with BSO and XAI.

A. Dataset Description

In the case of this research, we used the Campus Recruitment dataset found on Kaggle [27] that gives exhaustive data on the academic, demographic, and cognitive characteristics of students, as well as placement position and salary performance. The database consists of around 215 records, and the most important data points are 10 th and 12 th grades percentages, undergraduate performance, specialization, gender, previous work experience, and psychometric scores. The salary values in the dataset range approximately between 200,000 INR and 940,000 INR, with the majority of placed students receiving salaries within the 200,000–400,000 INR interval. Despite the absence of semester-wise temporal details within the dataset, the academic factors correspond to progressive milestones that signify the process of gradual academic development from high school education to university education. The milestones of progressive academic development form a structured sequence that is suitable for modelling academic evolution in relation to employment outcomes. These various features render the data ideal in the joint modeling of the placement classification and salary prediction problems in one analysis framework.

Despite the small number of records in the Campus Recruitment dataset of 215, there are several methodologies that have been adopted to prevent overfitting and enhance generalization. Firstly, the suggested methodology includes the use of Dropout regularization, weight decay, early stopping, and fivefold cross-validation to ensure stability when training the model on different datasets. Secondly, the use of a multi-branch model ensures the learning of complementary models instead of adding complexity by having parameters for the different branches. For instance, the MLP will be used to learn static cognitive skills, the BiLSTM learns the academic trajectory of the individual students, the Transformer learns interactions between contextual features, and finally, the GAT learns the dependencies between students. In addition to this, the use of Bolas Spider Optimization ensures that the model selects the optimal hyperparameters to avoid instability when training the model. The Campus Recruitment data set is used as a benchmark data set in the context of education for assessing the performance of the hybrid employability prediction models, but the size of this data set is limited in nature and might not adequately reflect the diversity found in larger university populations.

B. Preprocessing

Preprocessing of the data set included various important steps that guaranteed quality and consistency of data. The missing values of numerical features were treated with mean imputation, and the available values of categorical features were imputed with the mode of the respective column. Continuous variables, including grades and salary, were min max normalized to fall within the range of [0,1] was expressed in Eq. (1).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

One-hot encoding of nominal categorical features such as specialization and gender was used to facilitate the use of these features in models of neural networks. Also, psychometric scores and personality indicators were summarized into a cognitive trait vector $C \in R^d$ per student, which offers an economic representation of cognitive attributes of each student towards the predictive framework.

C. Framework Architecture

The proposed ExCogGNet model is a hybrid deep learning model that is a combination of MLP, BiLSTM, Transformer, and Graph Neural Networks (GNNs) with Bolas Spider Optimization (BSO) controlling the choice of hyperparameters and weights (Fig. 2). The design is cognitive-trait-aware

designed used to predict salaries, combining both the presence of static, sequential, relational representations, and allowing the explanation of the design by means of the XAI techniques. The architecture consists of several branches, which are trained to complement each other and optimize together on classification (placement) and regression (salary).

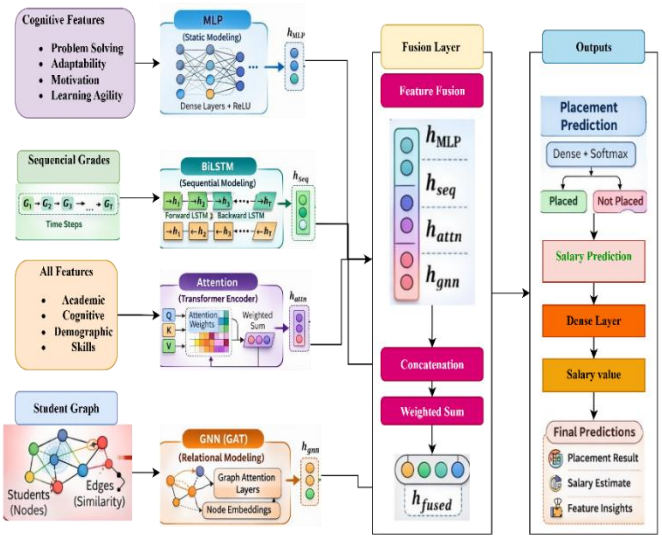


Fig. 2. Architecture of the proposed method.

1) MLP static cognitive modeling: The MLP branch takes the psychometric scores and personality indicators as they are extracted as vectors C to the MLP branch. The MLP is learned to find higher-order correlations between these features in the form of multi-layered hidden interactions. The hidden representation is calculated in Eq. (2).

$$h_{MLP} = f(W_1 C + b_1) \tag{2}$$

$f(.)$ represents the Relu Activation function, W1 represents the weight matrix, and b_1 represents the bias vector. The network uses numerous overlapping layers to simulate intricate patterns between cognitive characteristics that determine employability.

2) Sequential academic modeling using BiLSTM: To incorporate progressive academic development in the model for various levels of education, the model takes indicators like the percentage of secondary, higher secondary, and undergraduate education as the sequential development stage of education and is not based on a temporal sequence since there are no semester-based sequential data present in the dataset. However, the indicators show patterns that relate to academic growth and impact job prospects and salary outcomes in the future. Therefore, the BiLSTM model will be applied to capture the progression of academic development stages as sequenced in Eq. (3).

$$X_{seq} = [x_1, x_2, \ldots, x_T] \tag{3}$$

Such a sequence is processed by a Bidirectional LSTM (BiLSTM) both forward and backward, and therefore the model can make use of the context of the previous and upcoming semesters in Eq. (4).

$$h_{\overrightarrow{t}}, h_{\overleftarrow{t}} = LSTM(x_t, h_{\overrightarrow{t-1}}, h_{\overleftarrow{t+1}}) \tag{4}$$

Final sequential embedding the last forward and backward hidden states are concatenated to get the final sequential embedding in Eq. (5).

$$h_{seq} = [h_{\overrightarrow{T}}; h_{\overleftarrow{1}}] \tag{5}$$

The encoding of these trends in a temporal change in performance affects the placement likelihood and projected salary.

The rationale behind this proposed sequential modeling paradigm stems from contemporary approaches to incorporating self-attention mechanisms into the bidirectional LSTM architecture, which exhibited enhanced dependency modeling capabilities and training optimization stability through attention-based sequence modeling methodologies[15].

3) Contextual attention transformer encoder: To obtain inter-feature contextual interactions, the framework uses a Transformer encoder and multi-head self-attention. Embeddings of inputs are transformed into query (Q), key (K) and value (V) matrices, and the attention mechanism is calculated in Eq. (6).

$$Attention(Q, K, V) = softmax(\frac{QK^t}{\sqrt{d_k}})V \tag{6}$$

Multi-head attention enables the model to attend to multiple elements of the input at the same time so that it generates a contextual embedding h (trans) that augments the expression of intricate interactions among features.

4) Graph neural network for relational modeling: The students are modelled as nodes of a graph $G = (V, E)$ where the similarity is coded as an edge depending on the academic and cognitive characteristics. Specifically, two student nodes i and j are connected by an edge when cosine similarity exceeds a threshold of 0.85; the cosine similarity between the concatenated vectors of the normalized degree percentage, cognitive trait score, and technical skills score ensures that only meaningfully similar students are connected, yet the graph sparsity is maintained. Graph Attention Network (GAT) is an updated embedding of nodes based on the attentional message passing in Eq. (7).

$$h'_i = \sigma(\sum_{j \in N(i)} \alpha_{ij} W h_j) \tag{7}$$

The attention coefficients α_{ij} are determined using an adaptive calculation that is as follows in Eq. (8).

$$\alpha_{ij} = \frac{\exp\left(LeakyReLU(a^T[Wh_i \| Wh_j])\right)}{\sum k \in N(i) \exp\left(LeakyReLU(a^T[Wh_i \| Wh_k])\right)} \tag{8}$$

This process enables the network to give each student a different weight of peer influence, which reflects relational patterns that have an influence on placement and wages.

5) Bolas Spider Optimization (BSO) module: The BSO module is optimized on essential hyperparameters, such as

learning rate and hidden units, attention coefficients and fusion weights ($\lambda_{MLP}, \lambda_{BiLSTM}, \lambda_{transformer}, \lambda_{regression}, \lambda_{GNN}$). The fitness function is a combination of regression and classification goals in Eq. (9).

$$L_{total} = \lambda_1 L_{classification} + \lambda_2 L_{regression} \quad (9)$$

$L_{classification}$ is the cross-entropy loss when using to predict placement, and $L_{regression}$ mean squared error when predicting salary. BSO will perform a search through the search space, and the candidate solutions will be updated to reduce the total of L_{total}.

BSO was chosen in this experiment due to its adaptability in terms of achieving the optimal balance between explorative and exploitative searches when searching for optimal solutions within a heterogeneous hybrid architecture with multiple branches and attention mechanisms (Fig. 3). As compared to brute-force search techniques such as grid search and random search, BSO carries out optimization via population-based iterations on high-dimensional continuous parameter spaces. As compared to Bayesian optimization techniques that assume smooth surfaces of the objective function, BSO offers a flexible approach to global searching of the objective space in multi-task optimization scenarios.

Aside from hyperparameter tuning, the dynamic adjustment of the loss balancing coefficients λ_1 and λ_2 for multi-task is performed in the BSO model in order to reach a desirable balance between the accuracy in the classification of placement and the salary prediction. Candidate solutions are assessed during the optimization process in relation to both objectives, making it possible to have dynamic balancing.

6) Computational complexity analysis: The ExCogGNet's computational complexity comes from its joint computation of the MLP, BiLSTM, Transformer, and GAT branches. The computation needed by the MLP branch is O(ndh), where n refers to the number of samples, d is the dimensionality of features, and h stands for the number of hidden neurons. The BiLSTM branch brings sequential computation complexity of $O(Th^2)$, with T being the sequence length. The encoder of Transformer has a computation complexity of $O(T^2d)$ caused by self-attention mechanisms, while GAT necessitates O(|E|d) computation over graph edges E. Moreover, BSO optimization adds iterative search complexity of O(P×I), in which P is the size of the population, and I is optimization iterations. Even though the hybrid network model causes an increase in computation complexity compared to single networks, it offers much-improved predictive power, relational reasoning, and explanation.

7) Multi-task learning and explainable AI (XAI) combinations: The output of the feature embeddings of every branch is combined and sent to specialized multi-task output heads: a Softmax unit to classify the placements and a linear unit to predict the salary. SHAP values are calculated to provide both global feature importance and local prediction-level explanations, and attention weights of the Transformer and GNN parts are visualized to emphasize influential cognitive attributes and academic factors. This combined design guarantees the model provides not just a high predictive validity, but also clear and implementable information to career advisors and institutional decision-makers.

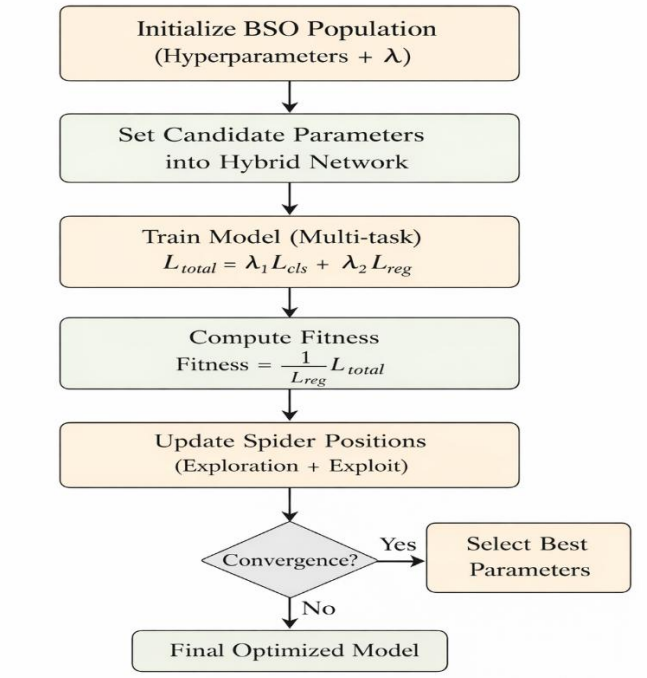


Fig. 3. Bolas Spider Optimization (BSO) workflow for hyperparameter and fusion weight optimization.

8) Counterfactual personalized intervention module: In addition to global and local SHAP accounts, the given framework has a counterfactual intervention module, which transforms SHAP feature attributions into action-oriented and student-specific suggestions. To every student forecasted as not placed, or forecasted below some salary level, the module finds the two most negative SHAP contributors and calculates how much the student needs to improve them in order to get his or her placement probability above the 0.5 mark or to get to his or her target percentile of salary. Consider, given that a student has a probability of placement, and the highest negative predictors are technical skills score and communication score, the module would give the particular level of improvement in each feature that is necessary to move the prediction to the placed category, giving career counselors an actual interpretable intervention strategy instead of an uncooked prediction.

The ExCogGNet algorithm proposed in (Algorithm 1) starts with pre-processing of the data such as normalization, encoding, and train-test splitting. Then a graph is created to depict the relationships among students in terms of similarity of features. The model combines the four branches: an MLP that captures cognitive features, a BiLSTM that captures temporal academic progression, a Transformer that captures contextual interactions, and a GNN that captures relationship learning. These embeddings are combined and sent to two output heads to perform classification in placement and salary prediction. The hyperparameters and fusion weights are tuned with Bolas Spider

Optimization and minimize a multi-task loss. Backpropagation is used to train the model in an iterative manner, and classification and regression measures are used to evaluate the model. Lastly, SHAP values and attention weights can be interpreted and make sense, thus allowing transparent and actionable decision-making.

Algorithm 1: BSO-Optimized Hybrid Network for Placement & Salary Prediction

```
Input:
  Dataset D with features (academic, demographic, cognitive)
  Labels: Placement (Y_cls), Salary (Y_reg)
Output:
  Trained model M
  Placement predictions Ŷ_cls
  Salary predictions Ŷ_reg
  Explainability insights (SHAP values, attention maps)
Begin
1. Data Preparation
  1.1 Load dataset D
  1.2 Handle missing values and encode categorical variables
  1.3 Normalize numerical features
  1.4 Split D into training set (D_train) and testing set (D_test)
2. Graph Construction
  2.1 Define nodes as students
  2.2 Compute similarity matrix based on feature space
  2.3 Construct adjacency matrix A
  2.4 Build graph G = (V, E)
3. Model Initialization (ExCogGNet)
  3.1 Initialize MLP branch for static cognitive features
  3.2 Initialize BiLSTM branch for temporal academic sequences
  3.3 Initialize Transformer branch for feature interactions
  3.4 Initialize GNN (GAT) branch for relational learning
  3.5 Initialize fusion layer
  3.6 Initialize output heads:
      - Softmax layer for placement classification
      - Linear layer for salary regression
4. Optimization using Bolas Spider Optimization (BSO)
  4.1 Initialize population of candidate solutions (hyperparameters)
  4.2 While stopping criterion not met:
      a. Evaluate fitness of each candidate:
          - Train model with given parameters
          - Compute multi-task loss:
              L = L_classification + λ * L_regression
      b. Update candidate solutions using BSO:
          - Exploration phase (lasso search)
          - Exploitation phase (reel refinement)
      c. Select best-performing configuration
  4.3 Set optimal parameters in model M
5. Model Training
  5.1 For each epoch:
      a. Forward pass:
          - MLP → static embeddings
          - BiLSTM → temporal embeddings
          - Transformer → contextual embeddings
          - GNN → relational embeddings
      b. Fuse embeddings
      c. Generate outputs:
          Ŷ_cls = Softmax(fused features)
          Ŷ_reg = Linear(fused features)
      d. Compute multi-task loss L
      e. Backpropagate and update weights
6. Model Evaluation
  6.1 Predict on D_test
  6.2 Compute metrics:
      - Accuracy, AUC (classification)
      - RMSE, R² (regression)
7. Explainability Module
  7.1 Compute SHAP values:
      - Global feature importance
      - Local explanations for each prediction
  7.2 Extract attention weights from:
      - Transformer layers
      - GNN layers
  7.3 Generate visualizations and insights
8. Output Results
  8.1 Return trained model M
  8.2 Return predictions (Ŷ_cls, Ŷ_reg)
  8.3 Return explainability outputs
End
```

IV. Result and Discussion

The Results and Discussion section considers the performance of the proposed framework, ExCogGNet, in terms of predicting student employability and salary, in terms of accuracy and interpretability. The performance of the model is evaluated based on the common classification and regression indicators, and compared with the traditional machine learning and state-of-the-art deep learning baselines. In addition to quantitative performance, the role of integration of cognitive traits, time progression, and relational learning in enhancing predictions is also examined in this section. Moreover, the results of explainability based on SHAP analysis and attention mechanisms are discussed to learn about some major factors affecting them. The results reveal the strength, generalization, and viability of the suggested method in the practical application of academic decisions in real-life situations.

A. Experimental Outcome

The general distribution of the Campus Recruitment data that was used in this study is shown in Table I. The sample size is 215 records of students; 148 students (68.8) were successful in placement, and 67 students (31.2) were not successful in placement. With regard to gender, the data sample is mostly male, with 139 students (64.7) representing it and 76 students (35.3) representing females. Concerning work experience, 65 students (30.2) have previous work experience, and 150 students (69.8) do not have previous work experience. The sample also contains two major specializations: 113 students (52.6) are in Commerce and Management, and 102 students (47.4) are in Science and Technology. In general, the table offers a concise summary of the demographic, academic, and placement-based features of the data, which illustrates that it is diverse and can be used to model not only the determinants of the placement outcomes but also the predictors of the salary.

Table II provides the detailed results of the optimization of the hyperparameters of the proposed BSO-based hybrid MLP-BiLSTM-Transformer-GNN model. In the case of the MLP branch, two hidden layers with 128 and 64 units, ReLU activation, and a dropout rate of 0.3 were used to reduce overfitting. The BiLSTM branch has 128 hidden units and a dropout rate of 0.2, and it is merely one layer and essentially captures the time-based dependencies in the sequential academic data. Transformer encoder is set with 4 attention heads, an

embedding dimension of 128, a feed-forward network (FFN) dimension of 256, and a dropout of 0.1, which enables the model to consider the contextual interactions between features. The GAT branch has 4 attention heads with a dimension of 128 as the output and ELU activation to learn relational embeddings over the student graph. All the branches are fused to the final branch by a weighted sum strategy where each branch's weight is 0.25. Training was done using the Adam optimizer, a batch size of 32, a learning rate of 0.0005, 100 epochs, and a weight decay of 0.0001.

TABLE I. DATA DISTRIBUTION TABLE

Category	Sub-Category	Count	Percentage (%)
Total Samples	—	215	100
Placement Status	Placed	148	68.8
	Not Placed	67	31.2
Gender	Male	139	64.7
	Female	76	35.3
Work Experience	Yes	65	30.2
	No	150	69.8
Specialization (Degree)	Commerce & Mgmt	113	52.6
	Science & Tech	102	47.4

TABLE II. OPTIMIZED HYPERPARAMETERS (BSO-BASED HYBRID MODEL)

Module	Key Settings
MLP	Layers: 2 • Units: (128, 64) • Activation: ReLU • Dropout: 0.3
BiLSTM	Layers: 1 • Hidden Units: 128 • Dropout: 0.2
Transformer	Heads: 4 • Embedding Dim: 128 • FFN Dim: 256 • Dropout: 0.1
GAT	Attention Heads: 4 • Output Dim: 128 • Activation: ELU
Fusion	Strategy: Weighted Sum • Weights (λ): 0.25, 0.25, 0.25, 0.25
Training	Optimizer: Adam • LR: 0.0005 • Batch Size: 32 • Epochs: 100 • Weight Decay: 0.0001

The distribution shows that most of the placements are within the ₹200k-₹400k range, which comprises almost 78.4% of the students placed (Fig. 4). Simultaneously, the existence of higher pay ranges indicates that there is an opportunity to have better pay among the high-achieving students, not only showing consistency in the overall placement results but also signifying the possibility of receiving high-paying jobs.

The Secondary percentage is 78.5%, with a standard deviation of 6.2, showing that the majority of the students were fairly consistent in performance, ranging between 60 and 95. The average of the Higher Secondary percentage is a bit higher at 81.2% and has less variability (5.8), indicating enhanced performance that is more consistent in senior secondary education, with a range of 65 to 98. The degree percentage in the undergraduate level is 75.6% with a wider variation (7.1), with values in the range of 55 to 95, meaning that students are more varied in academic results. Other than academic performance, the Cognitive Trait Score is 68.4 with a standard deviation of 8.5, showing moderate variations of cognitive abilities among the students. The highest variability is 9.2, with a mean of 72.3 of the Technological Skills Score, which indicates a great variation in technical competence with ranges of 50 to 90. On the whole, the table proves that there is a relative consistency in academic performance among students, but cognitive and technical skills are more diverse, and they are valued when it comes to distinguishing the general competencies and employability of the students.

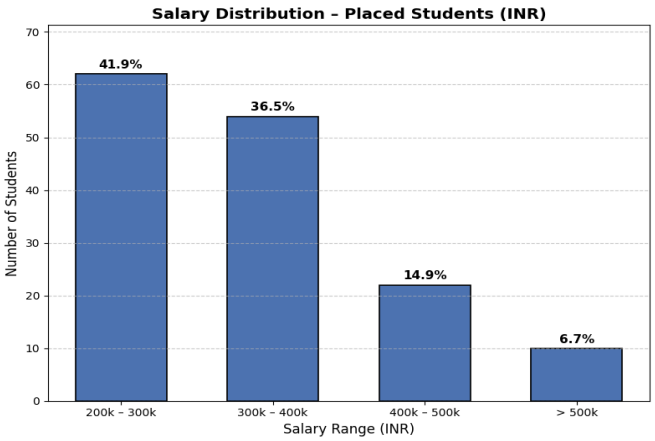


Fig. 4. Salary distribution.

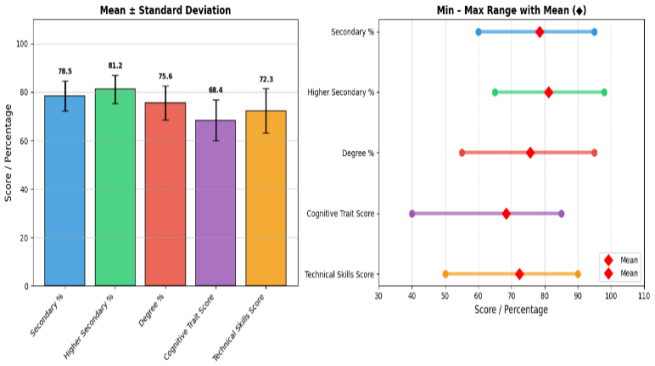


Fig. 5. Academic feature distribution.

The descriptive statistics in Fig. 5 show the mean performance of the students based on key academic, skill, and experience-related variables and the variability as well as the overall range of the scores. The Secondary Percentage (10 th grade) has a mean of 78.5 as well as a standard deviation of 6.2, as the performance of the students is consistent with scores ranging between 60% and 95%. The Higher Secondary %age (12 th) is a little higher, with the average being 81.2 % with a narrower spread (5.8), indicating a stronger and more consistent performance in senior secondary education. Degree Percentage, which is a measure of the performance of the undergraduates, has a mean of 75.6 and a higher standard deviation (7.1), with more varied academic results of 55-95.

It shows the feature contribution analysis with SHAP (SHapley Additive exPlanations), which measures the impact of each feature on the predictions of the model. The most significant features noted in the table, along with the importance score and the direction of the effect, are the Degree Percentage, with an importance score of 0.21. Its positive effect signifies that

the greater the undergraduate scores, the stronger the projected probability of achievement, as indicated in success, i.e., placement or anticipated performance. Such a strong positive influence of the Cognitive Trait Score (0.18) also highlights the importance of analytical and aptitude skills in the determination of results (Fig. 6).

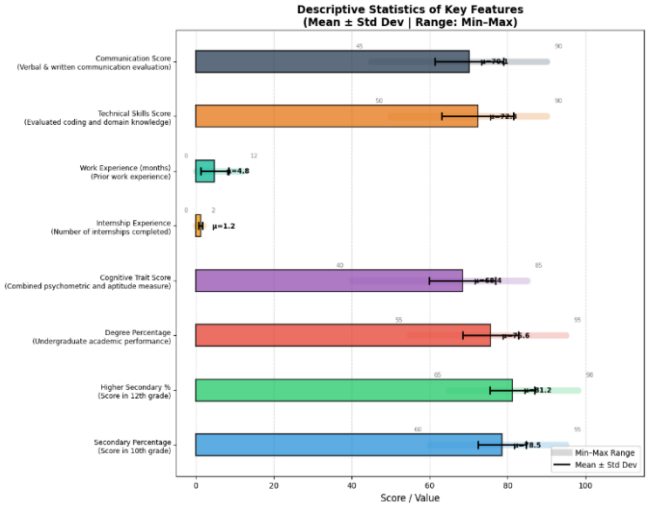


Fig. 6. Descriptive statistics of key features.

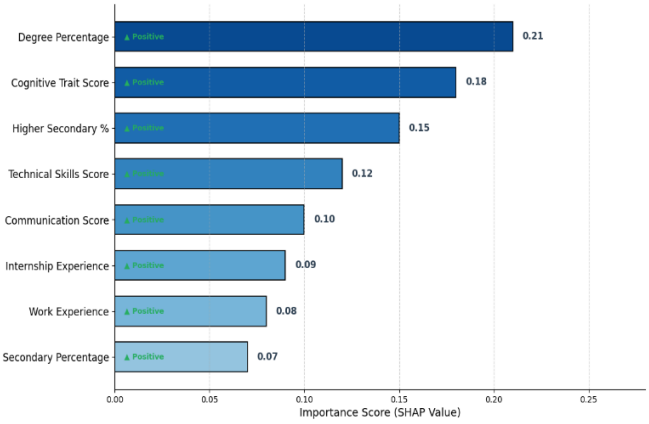


Fig. 7. Feature contribution analysis (SHAP-Based).

Fig. 7 shows the performance of the model based on standard evaluation measures: Precision, Recall, F1-score, and Support. In the Placed class, the model has a Precision of 97, meaning that most of the students that are predicted to be placed are actually placed correctly. There was a recall of 96%, indicating that the model was able to recognize 96% of the total students that were actually placed. The overall performance of this class is high, as the precision and recall balance; the F1-score is 96.5. The Placed class has a support of 106 students.

The confusion matrix helps to determine the effectiveness of the placement classification model in terms of comparison between the results predicted and actually obtained. Out of the 106 students who were actually placed, the model was correct in 102 of them being placed, and 4 students were misclassified as not placed. Of the 59 students who had not been placed, 56 of them were correctly identified as not placed, and 3 of them were not identified as placed. These findings show that the model is very accurate and the misclassifications in both classes are very minimal. On the whole, the confusion matrix illustrates that the model is very useful in differentiating placed and not placed students, which means that it is very reliable in making predictions of placement results (Fig. 8).

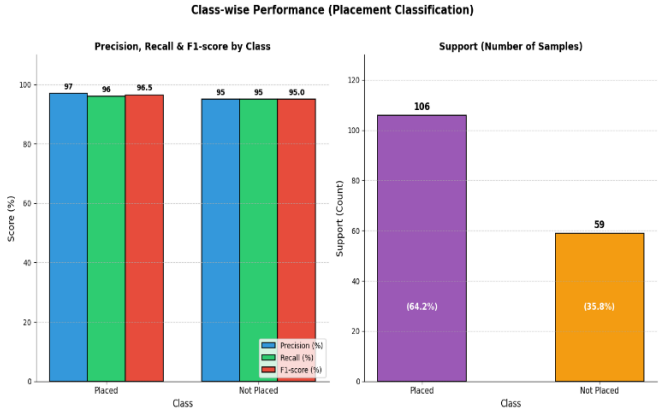


Fig. 8. Class-wise performance.

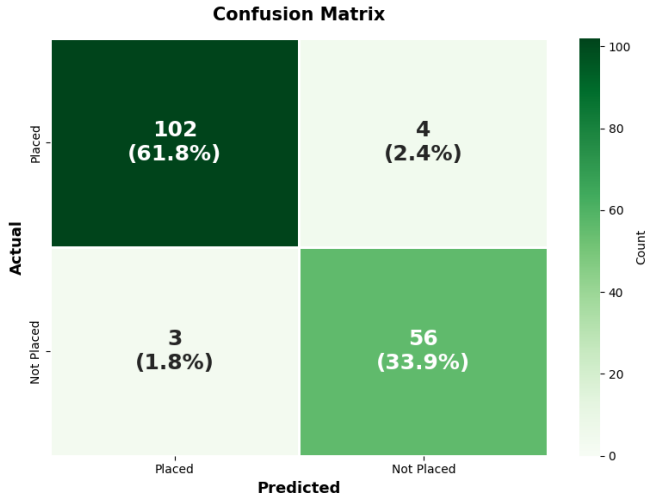


Fig. 9. Confusion matrix.

Fig. 9 gives the performance of different machine learning and deep learning models in terms of placement prediction in various evaluation metrics, such as accuracy, precision, recall, F1-score, and AUC. Conventional models such as SVM had moderate performance with an accuracy of 78% and AUC of 0.81, which reflected decent but poor predictive powers. Random Forest and XGBoost enhanced the performance, both achieving 82% and 84 % of the accuracies, respectively, with correspondingly higher rates of precision, recall, F1-score, and AUC, indicating more aptitude to address the complicated interaction of features.

It shows the salary prediction performance of various regression models in terms of RMSE, MAE, and R 2. Linear regression has the poorest results with an RMSE of 48k INR, an MAE of 36k INR, and R^2 of 0.62, which suggests that it has very poor predictive power. Random Forest and XGBoost ensemble models bring the predictions to RMSE of 42k and 38k INR, MAE of 31k and 28k INR, and R^2 of 0.71 and 0.75, respectively,

showing more non-linear relationships in the data. Deep learning architectures such as BiLSTM and Transformer are also used to improve the performance, with RMSE of 35k and 33k INR, MAE of 25k and 23k INR, and R^2 of 0.78 and 0.80 indicating capability to capture sequential and complex feature interactions. The GNN, which does not use BSO, exhibits further improvement with an RMSE of 31k INR, MAE of 21k INR, and R^2 of 0.82, which indicates the advantage of capturing relational information in the model. Compared to all other models, the proposed hybrid model of MLP, BiLSTM, Transformer, and GNN with BSO optimization has yielded data with an RMSE of 24k INR, MAE of 18k INR, and R^2 of 0.91, which includes highly predictive and reliable data in terms of accuracy and performance of the predicates in the salaries. Given that the salary range in the dataset falls between roughly 200k INR and 940k INR, an RMSE of 24k INR can be regarded as quite low, and hence a sign of the reliability of the regression model (Fig. 10).

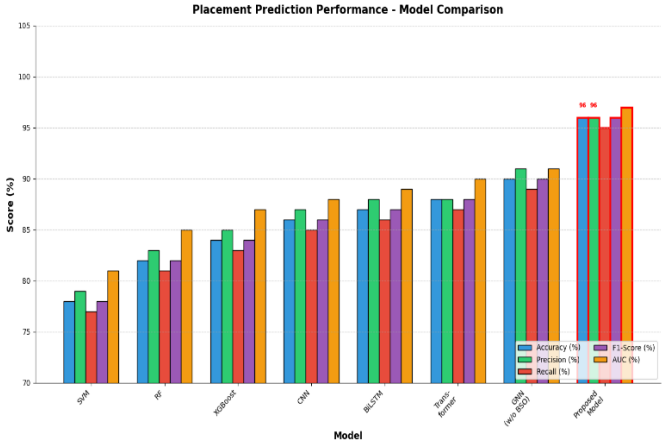


Fig. 10. Placement prediction performance.

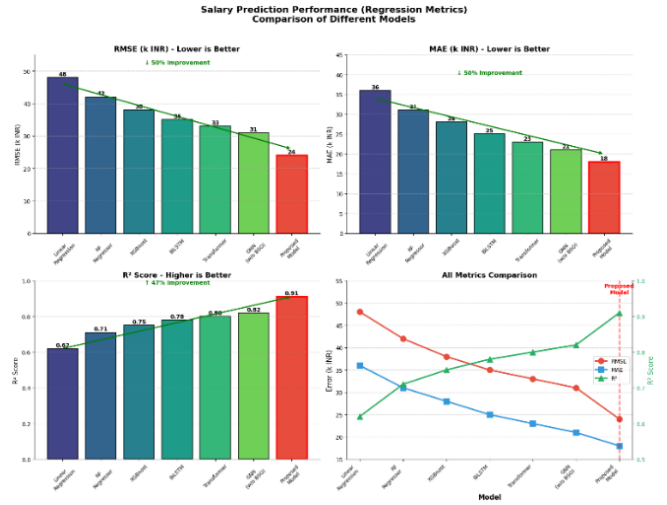


Fig. 11. Salary prediction performance.

B. Performance Analysis

Fig. 11 illustrates how the model accuracy of training and testing accuracy improved with 100 epochs. At the first 10 epochs, the training accuracy is 88.5, and the testing accuracy is 86.2, which means that the model is in the process of learning but still can be improved. The higher the number of epochs, the higher the training and testing accuracy, indicating that there is an efficient learning and generalization. At 50 epochs, the training and testing accuracy are 95.5 and 94.0%, respectively, which is good model performance with very little overfitting. After 60 epochs, the accuracy gains start to level off, and the training accuracy reaches 97.5 per cent, whereas the testing accuracy reaches around 96.0 per cent. In general, the table proves that the model is convergent, as its accuracy is high and stable on the training and testing data, which proves strong learning and strong generalization on the unknown data (Fig. 12).

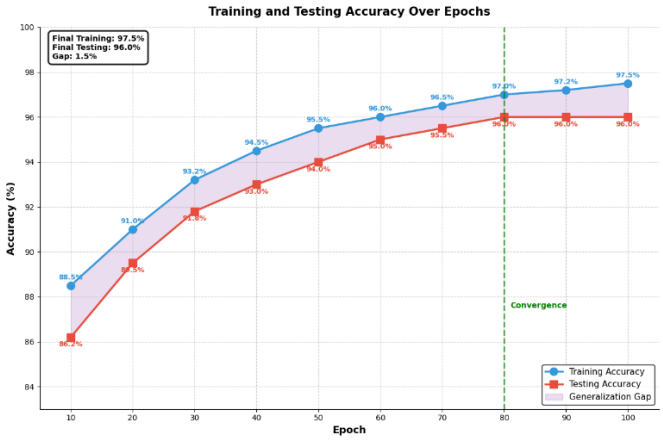


Fig. 12. Training and testing accuracy table.

The plot of the training and testing loss of the model over 100 epochs demonstrates the capability of the model to learn and generalize well. The training and testing losses at age 10 are 0.42 and 0.45, respectively, which means that the model is yet to learn much. Training loss and testing loss smoothly decrease as training advances, indicating improved model performance and enhanced precision in prediction. Training loss decreases to 0.18, and test loss decreases to 0.21 by 50 epochs, indicating that the model is learning the data patterns and still has good generalization. The loss keeps decreasing more slowly after 80 epochs, with 100 epochs having the results of 0.06 (training) and 0.10 (testing). In general, the table shows that there has been steady convergence of the model, with both training and testing loss reducing continuously, which is a sign of effective learning and very little overfitting (Fig. 13).

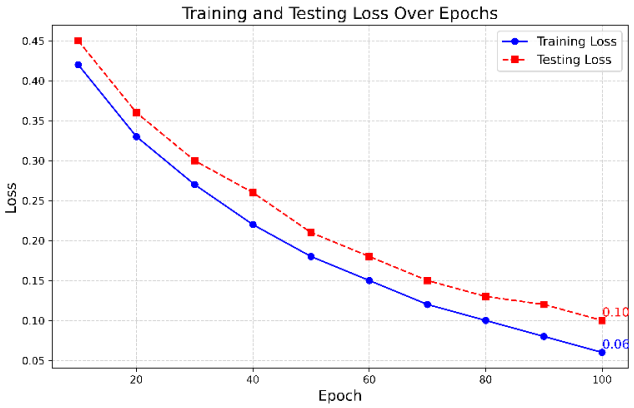


Fig. 13. Training and testing loss.

The convergence and optimization stability of the process is depicted in the graph after a number of iterations (Fig. 14). It presents three metrics, which are BSO fitness, placement accuracy, and salary R 2, versus the number of iterations. The three metrics have a slow increase with increment of the iterations, which demonstrates that the optimization algorithm is indeed refining the solution. The BSO fitness is steadily increasing between 0.88 and 0.96 with a steady increase in the objective function. The same case is with the accuracy of placements, whereby the accuracy rises to 96 % compared to 90 %, indicating that the model is more predictive of successful placements as time goes by. Salary R^2 also increases to 0.91 as compared to 0.84, which indicates a better correlation between predicted and actual salaries. The convergence of the algorithm is also identified in the graph and the stability of the model as the measures reach a steady level in the later iterations, indicating that additional iterations do not significantly alter them.

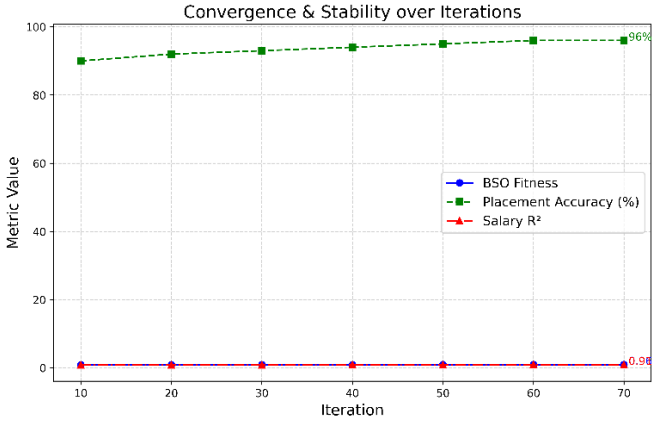


Fig. 14. Convergence & stability.

C. Ablation Study and Computational Justification

An ablation study has been performed by successively removing each component of the hybrid architecture and observing the impact on the performance of ExCogGNet. The assessment was made based on both placement classification accuracy and the salary R^2 performance. Experimental results have shown that the overall ExCogGNet system has the best performance with placement classification accuracy of 96%, and a salary R^2 score of 0.91 (Table III). Removing the Transformer branch leads to a loss of ability for contextual interactions in learning. Removal of the GNN branch decreases relational dependency analysis skills among students. The removal of the BiLSTM branch affects ordered academic progressions in the learning process. Lastly, the exclusion of the MLP branch limits cognitive feature representation skills. It can be seen that all modules provide complementary skills to the whole system; hence, the improvements in the performance cannot be caused by excessive architectural complexity. In addition, even though using the hybrid architecture is computationally more expensive than using standalone models, the huge boost in the accuracy and robustness of multi-tasking makes this extra complexity worthwhile.

D. Multi-Task Loss Weight Sensitivity Analysis

In order to measure the stability of the multi-objective optimization, further experimental studies were performed using variations in the loss balancing parameters λ_1 and λ_2 for classification of placements and salary regression, respectively. It can be observed from experimental data that an increase in λ_1 results in better classification quality but increases salary prediction RMSE, and an increase in λ_2 improves regression results but adversely affects classification results. The optimal results in terms of the trade-off between both objective functions have been found when $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$. Also, the small value of the performance deviation across different optimization iterations is evidence of the algorithm's stability and consistency, as shown in Table V.

TABLE III. ABLATION STUDY OF EXCOGGNET COMPONENTS

Model Variant	Placement Accuracy (%)	Salary R^2
Full ExCogGNet (MLP + BiLSTM + Transformer + GNN + BSO)	96	0.91
Without Transformer	93	0.86
Without GNN	91	0.84
Without BiLSTM	89	0.82
Without MLP	88	0.80
Transformer Only	86	0.78
BiLSTM Only	84	0.75
MLP Only	81	0.71

TABLE IV. EFFECT OF LOSS WEIGHT VARIATIONS ON MULTI-TASK PERFORMANCE

λ_1 (Classification)	λ_2 (Regression)	Placement Accuracy (%)	Salary RMSE (INR)
0.2	0.8	88	22k
0.4	0.6	92	23k
0.5	0.5	94	25k
0.6	0.4	96	24k
0.8	0.2	97	29k

Fig. 15 shows the cross-validation 5-fold performance of the classification model. It is a summary of training and validation accuracy, precision, recall, F1-score, and AUC of five folds, the average, and standard deviation of these values. The findings show that there is a high consistency level amongst folds and that the average accuracy of training is 97.1 % and average accuracy of validation is 96.0 %. Precision and recall rates are also good, which means that the model has good capabilities of recognizing true positives and reducing falsely recognized positives and falsely recognized negatives. The overall classification performance is confirmed by the F1-score that balances recall and precision and has an average of 96.0. The solid averaged values of the AUC, 0.97, also contribute to the high discriminative ability of the model. Low standard deviations in all the metrics indicate that the performance was consistent and variable was small, and it shows the reliability of the model and its generalizability in various subsets of the data.

The minimal standard deviation seen in the five validations implies that there is consistent learning performance and less sensitivity to changes in training data due to the moderate number of samples used. This implies that the proposed framework is able to learn generalized features rather than learning specific characteristics of the dataset. It is, however, important to conduct further testing on a large, multisite database comprising students of varied backgrounds.

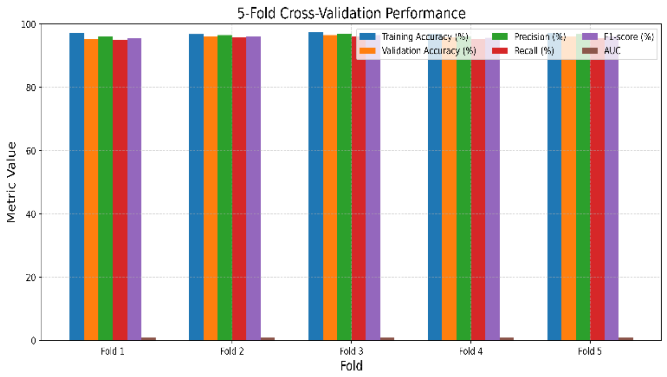


Fig. 15. 5-Fold cross-validation performance (classification).

The graph (Fig. 16) indicates the performance of the salary regression model when cross-validation is conducted five-fold. It shows three metrics, namely RMSE, MAE, and R^2 against each fold. RMSE reduces steadily with each fold, with fold 1 of 26 k INR in fold 5 of 24 k INR, and MAE also reduces to 18 k INR, showing that the prediction errors are reduced, as well as the model becomes more accurate across folds. The values of R^2 are quite high, between 0.89 and 0.92, and this shows a high correlation between the predicted and actual salaries. The graph shows the reliability and stability of the regression model since the measures indicate minimal change in different folds, which proves the ability of the model to generalize.

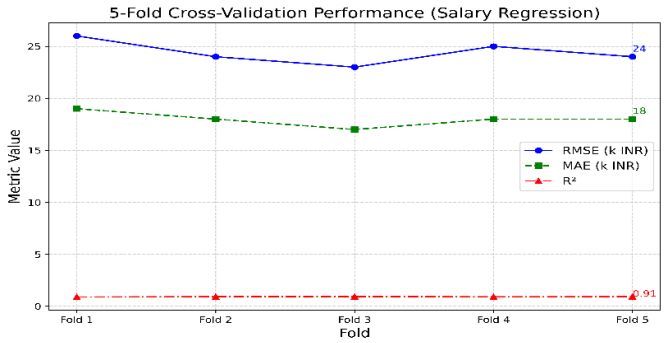


Fig. 16. 5-Fold cross-validation performance (regression).

The comparison table on the baseline gives the performance of several conventional machine learning and deep learning models in comparison with the proposed hybrid MLP-BiLSTM-Transformer-GNN model trained using Bolas Spider Optimization as a placement and salary predictor. The findings show that the classical models, i.e., SVM and random forest, are moderate in terms of placement accuracy of 78 % and 82 %, respectively, and relatively higher in terms of errors in salary prediction, represented by an RMSE value of 48k INR and R^2 values of 0.62 and 0.71, respectively. CNN, BiLSTM, and Transformer models are advanced deep learning models that achieve better performance, with placement accuracies ranging between 86% and 88% and decreasing lower values at RMSE, proving the benefit of sequential and contextual learning. Graph neural networks with no BSO can achieve 90 % placement accuracy, which shows the effectiveness of relational modeling. The hybrid model that is proposed and includes BSO is in a much better position than any of the baselines, with 96% placement accuracy, 96% F1-score, and 0.91 predictive R^2 of salary.

Regarding the evaluation of the effectiveness of the suggested framework under small data situations, five-fold cross-validation techniques were applied, together with regularization-based optimization approaches within the experimentation stage. As a result, the hybrid system consistently performed at a high level of validation with low variability between different folds. This indicates stable learning behavior and insensitivity to variations in the training data set. The results indicate that the model learned generalized representations of features instead of being sensitive to specific aspects of the used data set, despite its relatively small size. However, the selected data set is a benchmark educational data set based on the Campus Recruitment data set, including around 215 records. Hence, it does not necessarily represent multi-faculty environments adequately due to its scale. Thus, while reporting a 96% placement success rate is a good benchmark result, further large-scale validation is required.

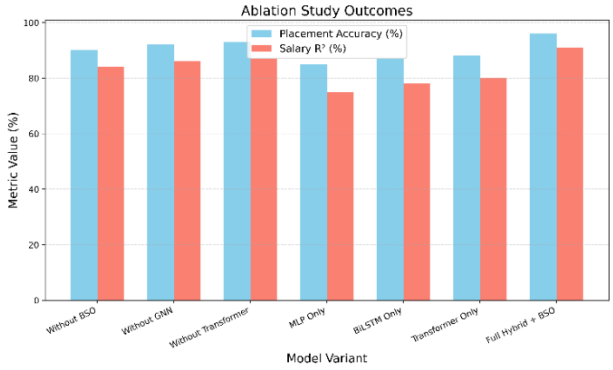


Fig. 17. Component contribution analysis.

The ablation study analysis, as shown in Table IV and Fig. 17, shows how each component individually contributes to the overall predictive performance of ExCogGNet. The complete hybrid model optimized through BSO resulted in maximum placement accuracy at 96% and maximum salary prediction at 0.91. The omission of BSO resulted in a decrement in optimization stability as well as a performance drop, thus emphasizing the importance of optimization of hyperparameters and fusion weights. Omission of the GAT branch resulted in weak relational learning in the case of students as well as prediction performance, thus underlining the importance of modeling peer relationships through graph learning. The exclusion of the Transformer resulted in the inability to learn from contextual feature interactions, while excluding the BiLSTM branch led to inadequate modeling of academic progress over time for students. Moreover, the standalone performance of MLP was significantly low since only static

features can predict employability. This analysis proves the need for a hybrid multi-branch architecture for effective multi-tasking.

To confirm that the performance difference between the proposed model and the strongest base, GNN without BSO, is statistically significant and cannot be explained by random fluctuation, the McNemar test was used to test the placement classification results of the five cross-validation folds. The p-value of 0.003 confirms that the increase in accuracy between 90% and 96% is statistically significant at the level of significance of 0.01, proving that the gains are real and reproducible and not data-specific.

TABLE V. BASELINE COMPARISON FOR PLACEMENT AND SALARY PREDICTION

Model	Placement Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	RMSE (k INR)	MAE (k INR)	R^2
SVM	78	79	77	78	0.81	48	36	0.62
Random Forest (RF)	82	83	81	82	0.85	42	31	0.71
XGBoost	84	85	83	84	0.87	38	28	0.75
CNN	86	87	85	86	0.88	35	25	0.78
BiLSTM	87	88	86	87	0.89	33	23	0.80
Transformer	88	88	87	88	0.90	31	21	0.82
GNN without BSO	90	91	89	90	0.91	29	20	0.84
Proposed Hybrid MLP-BiLSTM-Transformer-GNN + BSO	96	96	95	96	0.97	24	18	

E. Discussion

The suggested hybrid MLP-BiLSTM-Transformer-GNN model, optimized with Bolas Spider Optimization (BSO), provided an accuracy of 96% in the placement of the model, a 0.97 AUC, and a 0.91 salary prediction R-squared. These scores beat any of the baseline models. Furthermore, when interpreted relative to the overall salary distribution of the dataset, the observed RMSE indicates that the prediction deviation remains comparatively small, supporting the practical usefulness of the proposed framework for salary estimation tasks. such as GNN (without BSO) and Transformer (without BSO) (90 and 88 percent accuracy, respectively). The statistical significance of such an improvement, proved with the help of the McNemar test ($p = 0.003$), is the fact that the performance improvements are not specific to the dataset. The ablation analysis shows that BSO is essential for the globally optimal fusion weights and attention coefficients of the heterogeneous branches. The loss of BSO caused a significant decrease in performance, which proves that when the architecture is simply combined without any principled optimization, it may produce suboptimal learning. Even though BSO has been shown to exhibit stable convergence behavior along with robust predictive performance under the proposed framework, it is noteworthy that this research effort fails to provide an experimental benchmarking analysis of the optimization performance of BSO compared to other traditional approaches such as Bayesian optimization, grid search, and random search. The sensitivity analysis on the loss functions of the multi-task model shows that the proposed optimization framework is quite robust. From experimental results, it was found that varying settings of λ_1 and λ_2 had an effect on the balance between classification performance in placement and regression in salary prediction. The chosen setting led to convergence and high performance in both classification and regression. In addition, the repeated observation of the optimization process over several iterations shows the stability of BSO-based adaptive optimization.

The SHAP-based analysis revealed that the degree percentage (0.21), cognitive trait score (0.18), and technical skills score (0.12) are the most significant predictors, which supports the empirical argument of the cognitive-aware design of the model. The counterfactual SHAP explainability module offers feature-level insights into attribution and creates recommendation guidance for users by leveraging influential academic and cognitive attributes. These explanations can help career counselors and other institutional representatives gain insights into the significant determinants that influence employment results. However, in this particular research, no human-in-the-loop evaluation was conducted, nor did deployment testing or validation with career counselors or educators take place. Thus, this recommendation tool is more of a prototype that demonstrates how the proposed solution works through explainable means. The next step in future research is human-centered testing, usability evaluation, and institutional deployment research. Even though excellent prediction has been made through the use of the ExCogGNet, one challenge that should be taken note of is the limitation brought about by the size of the datasets used in the study. For deep hybrid networks, it is critical to use diversified datasets that help in increasing the generalization ability of the architecture due to its heterogeneity within different educational settings. The problem has been resolved through the use of five-fold cross-validation, dropout, optimized weights, and multi-tasking techniques to avoid overfitting. Furthermore, the constant accuracy levels obtained in validation, despite the small number of datasets, suggest stable learning behavior. Nevertheless, further validation using more diversified datasets from different universities will be needed.

V. CONCLUSION AND FUTURE WORK

In this study, an Explainable Cognitive Graph Network, ExCogGNet, was proposed that integrates MLP-based encoding of cognitive traits in a static form, BiLSTM-based modeling of ordered academic progression stages, Transformer-based contextual feature attention, and Graph Attention Network learning of peer-level relational structures within one multi-task

model. The model also does placement classification and salary prediction, and Bolas Spider Optimization across the world optimizes hyperparameters, attention coefficients, and fusion weights, with a combined cross-entropy and mean squared error loss, alleviating heterogeneous architecture instability.

The experimental findings on the Campus Recruitment benchmark dataset demonstrate that ExCogGNet is 96% accurate in placement and has an AUC of 0.97 and a salary prediction RMSE of 24k INR and R2 of 0.91, which is much higher than baseline models (SVM 78%), Transformer (88%), and un-optimized GNN (90%). The improvement is statistically significant ($p = 0.003$), as proven by McNemar. One important contribution is the counterfactual SHAP-based explainability module, which generates interpretable, recommendation-oriented insights for potential use in educational decision-support environments, converting feature attributions into individual, actionable recommendations on student skill improvement, which may support future decision-making processes for career counselors and institutional stakeholders after further practical validation. The ablation study results further confirm that each architectural component, including MLP, BiLSTM, Transformer, GNN, and BSO optimization, contributes complementary predictive capability, thereby justifying the hybrid multi-branch design of ExCogGNet. The ability of the framework to generate individualized recommendation-oriented explanations enhances its potential applicability in academic advisory systems, placement preparation programs, and institutional career guidance environments.

Though the introduced ExCogGNet framework exhibits good prediction performance on the Campus Recruitment dataset, its modest sample size makes comprehensive evaluation of its generalization ability challenging across a wide range of different educational organizations. As such, efforts will be directed at testing the framework on bigger educational datasets that include diverse student populations, behavioral learning history, socioeconomic features, and current academic progression information from multiple institutions. Despite the promising results obtained in terms of predictive performance and explainability, the present study has not considered user-focused validation through experiments with actual users such as students and career counselors. Therefore, the recommendatory explainability component should be seen as a prototype model for now. Usability experiments, human expert analysis, and real-world application will be conducted in subsequent research endeavors. Other potential research directions include methods of constructing graphs in scalable ways, employing light transformers, and utilizing federated learning techniques.

In the current dataset, milestone-based educational progression is included compared to fine-grained longitudinal academic history; future research might also include semester-based academic progress for improving sequential learning ability.

REFERENCES

[1] M. Vivyn, T. Rani, and others, "Comprehensive Career Placement Predictor: An Analytical Tool for Optimizing Job Placement Outcomes," in 2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), IEEE, 2024, pp. 1–6.

[2] K. Khalid and M. A. K. Al Zahrani, "Investigating the dynamics of workforce stability: The influence of perceived employability on career resilience, task performance, learning agility and employee retention," Eur. J. Sustain. Dev., vol. 13, no. 3, pp. 389–389, 2024.

[3] S. A. K. Malalha, M. Burhanuddin, and N. B. M. Yunos, "Unveiling the tapestry of machine learning: A comparative analysis of support vector machines random forests and neural networks in diverse applications," Tuijin JishuJournal Propuls. Technol., vol. 45, no. 3, p. 2024, 2024.

[4] K. W. Manikrao, S. Allagi, W. Y. Leong, and M. Laddi, "An Explainable AI for Stock Market Prediction: A Machine Learning Approach with XAI and Deep Neural Networks," J. Comput. Cogn. Eng., 2025.

[5] Y. Yang and Y. Yao, "Graph Neural Network–Driven Modeling of Student Relationships and Optimization of Activity Organization.," Rev. Comput. Eng. Stud., vol. 12, no. 3, 2025.

[6] S. F. Ahmed et al., "Enhancement of traffic forecasting through graph neural network-based information fusion techniques," Inf. Fusion, vol. 110, p. 102466, 2024.

[7] H. Qawaqneh et al., "Bolas Spider Algorithm: A Novel Efficient Nature-Inspired Metaheuristic for Complex Continuous Optimization.," Int. J. Intell. Eng. Syst., vol. 19, no. 1, 2026.

[8] R. N. Alibayeva and S. U. Allayarova, "From Data to Pedagogy: Leveraging Explainable Artificial Intelligence to Enhance Trust, Transparency, and Effectiveness in Intelligent Learning Systems," Qubahan Techno J., vol. 4, no. 3, pp. 51–65, 2025.

[9] C. Poblete and A. Aktas, "The role of conscious awareness in the relationship between entrepreneurs' hubris and the degree of internationalization," Cogent Bus. Manag., vol. 12, no. 1, p. 2560651, 2025.

[10] N. T. Mai, W. Cao, and W. Liu, "Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data," Appl. Sci., vol. 15, no. 17, p. 9605, 2025.

[11] Q. Huang and Y. Zeng, "Improving academic performance predictions with dual graph neural networks," Complex Intell. Syst., vol. 10, no. 3, pp. 3557–3575, 2024.

[12] V. P. Dwivedi and X. Bresson, "A Generalization of Transformer Networks to Graphs," 2020, arXiv. doi: 10.48550/ARXIV.2012.09699.

[13] E. Ofori and D. K. Dake, "Explainable artificial intelligence in LSTM transformer models for student performance analysis," Discov. Comput., vol. 28, no. 1, p. 313, Dec. 2025, doi: 10.1007/s10791-025-09814-9.

[14] H. Aljuaid, A. Alhothali, O. Al-Zamzami, and H. Assalahi, "TransGAT: Transformer-Based Graph Neural Networks for Multi-Dimensional Automated Essay Scoring," ArXiv Prepr. ArXiv250901640, 2025.

[15] S. Dontu, S. R. Addula, P. K. Pareek, R. Vallabhaneni, and M. M. Adnan, "Attack detection from Internet of Things using TPE based self-attention based bidirectional long-short term memory," in 2024 international conference on intelligent algorithms for computational intelligence systems (IACIS), IEEE, 2024, pp. 1–6.

[16] G. Vijh, P. P. Singh, S. Tiwari, S. Vijh, and S. Kumar, "Predictive Modelling of Salary Trends in the Tech Industry Using Machine Learning Algorithms," in The International Conference on Recent Innovations in Computing, Springer, 2024, pp. 38–47.

[17] L. Wang, "Deep Learning for College Graduates Employment Prediction: A Computational Approach," IEEE Access, 2026.

[18] A. R. Alsheyab, M. Alkhasawneh, and N. Shahin, "Job Market Cheat Codes: Prototyping Salary Prediction and Job Grouping with Synthetic Job Listings," 2025, arXiv. doi: 10.48550/ARXIV.2506.15879.

[19] M. Xu, "Salary prediction using machine learning," Sch. Rev. J., vol. Summer 2025, no. 13, Jun. 2025, doi: 10.70121/001c.139043.

[20] A. BISWAS, "A Comparative Study for Salary Prediction Based on Different Models of Machine Learning," 2023.

[21] Habibu Aminu, Badamasi Imam Yau, Fatima Umar Zambuk, Emanuel Ramsom Nanin, Abdulmutallib Abdullahi, and Ismail Zahradeen Yakubu, "Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm," Jan. 2024, doi: 10.5281/ZENODO.10629245.

[22] A. M. Ahmed, T. A. Rashid, and S. Ab. M. Saeed, "Cat Swarm Optimization Algorithm -- A Survey and Performance Evaluation," 2020, doi: 10.48550/ARXIV.2001.11822.

[23] Q. Cappart, D. Chételat, E. Khalil, A. Lodi, C. Morris, and P. Veličković, "Combinatorial optimization and reasoning with graph neural networks," 2021, doi: 10.48550/ARXIV.2102.09544.

[24] M. M. Sesay, A. Ngunyi, and H. Imboga, "An explainable hybrid transformer–BiLSTM framework for multivariate household energy demand forecasting with weather integration," Energy Inform., Feb. 2026, doi: 10.1186/s42162-026-00635-8.

[25] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," Sci. Rep., vol. 14, no. 1, p. 24882, Oct. 2024, doi: 10.1038/s41598-024-76079-5.

[26] M. Cedro and D. Martens, "GraphXAIN: narratives to explain graph neural networks," in World Conference on Explainable Artificial Intelligence, Springer, 2025, pp. 91–114.

[27] Ben Roshan, "Campus Recruitment." 2020. [Online]. Available: https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement

Large Language Models for Arabic Automated Essay Scoring

Leena Najjar, Liyakathunisa Syed
Department of Computer Science-College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia

***Abstract*—Automated Essay Scoring (AES) has become an important research area in educational artificial intelligence due to its potential to support scalable and consistent assessment. The developments within the realm of transformers and large language models (LLMs) have led to great improvements within AESs through their ability to comprehend semantics and context, as well as their knowledge of rubrics. Although these advancements have been realized within the English language, there is still relatively little research surrounding Arabic Automated Essay Scoring (AAES). This survey summarizes some of the latest advancements in AAES and discusses traditional ML models, deep learning, transformers, and LLM-driven evaluation frameworks. In this study, researchers synthesize the relevant literature regarding datasets, prompts, pre-processing methods, performance metrics, and reliability of scores. Typical performance metrics used to analyze the level of agreement between human raters and automated systems are QWK, MAE, and correlation-based metrics. The survey also describes crucial challenges encountered by AAES systems such as insufficient amount of data, inconsistencies, high computation costs, bias, and non-reproducibility. Overall, it can be said that both transformers and LLMs achieve better performance when it comes to capturing context information and providing agreement with human assessment. However, issues with reproducibility and scalability continue to persist. Additionally, the survey presents new areas of research that may be relevant to future AAES studies, such as multilingual evaluation, hybrid grading, explainability, and standardized sources of Arabic essays.**

Keywords*—*Arabic automated essay scoring; large language models; educational AI; prompt engineering; evaluation metrics; natural language processing

I. Introduction

Artificial intelligence is transforming how today's educational systems measure student outcomes, covering aspects ranging from feedback mechanisms to test construction and grading systems [1], [2]. Within this emerging framework, Automated Essay Scoring (AES) is now an important focus of ongoing research, designed to reduce teachers' burdens while increasing consistency, objectivity, and scalability in grading. The earlier Automated Essay Scoring (AES) systems relied on machine learning approaches, which in turn depended on feature engineering. These earlier systems were extremely efficient in the prediction of scores, yet failed to grasp the essence of the essays, the underlying structure of the arguments, and their context [3], [4].

The emergence of deep learning and transformer-based models is an important milestone in the development of automated grading. These models now utilize context and long-range dependencies in essays and other forms of text to better understand language. For instance, Siamese Bi-LSTM models and transformer-based Automated Essay Scoring tools are becoming closer to human judgment than earlier machine-learning-based systems because of their better semantic understanding of the content of the essays [5], [6]. Recently, large language models (LLMs) have moved the AES paradigm towards reasoning-oriented evaluation models that utilize models of natural language interaction, such as rubric-informed prompting, few-shot learning, and instruction tuning to mimic human grading behavior [7], [8].

AAES systems also have their challenges, which are largely due to the complexities of the Arabic language and the lack of sufficiently large datasets for training and evaluating these systems. For instance, although multilingual large language models are highly promising in many evaluation scenarios, their application in Arabic Automated Essay Scoring systems is relatively less explored [8], [9]. In addition, there is a lack of comprehensive systematic studies on the evolution of these systems, dataset characteristics, and evaluation frameworks in the context of Arabic Automated Essay Scoring systems in the existing body of research.

To address this identified information gap, this study aims to synthesize existing research on Arabic automated essay scoring powered by large language models. It discusses the evolution of architectures, the creation of Arabic evaluation datasets, prompt design, the measurement of results, and the ethical issues related to the deployment of AI technology in education. With the synthesis of information from multiple research threads, the study aims to provide a clear reference guide for the development of Arabic automated essay scoring technology that is effective, scalable, and suitable for educational contexts.

The study makes the following contributions to the field of Arabic automated essay scoring powered by large language models: it proposes a comprehensive architectural taxonomy for AAES, presents a discussion on the creation of Arabic evaluation resources, and compares prompting and evaluation techniques for LLM-based essay grading. To the best of our knowledge, this is one of the first systematic surveys on Arabic Automated Essay Scoring (AAES) powered by large language models.

Unlike other surveys on automated essay scoring, this study provides a comprehensive view of Arabic automated essay scoring powered by large language models by discussing together the evolution of architectures, the creation of Arabic evaluation resources, the designing prompts, and the performance evaluation.

The paper is divided into the following sections: Section II discusses the systematic review methodology, Section III synthesizes the findings, Section IV discusses the implications and challenges, and Section V concludes with the future research directions.

II. Methodology

This survey is conducted using a PRISMA-based methodology to maintain transparency, reproducibility, and comparability of the results with other studies on Arabic Automated Essay Scoring (AAES). The relevant papers were retrieved from widely recognized scientific data sources, including IEEE Xplore, ScienceDirect, SpringerLink, MDPI, ACM Digital Library, and arXiv. The scope of the study included papers published between 2018 and 2026, focusing on automated essay scoring, Arabic NLP, large language models, and AI-based educational evaluation.

A four-step paper selection methodology was employed as follows:

- Identification: The relevant papers were identified with the help of the search terms automated essay scoring, Arabic NLP, transformer-based models, and large-language-model-based evaluators.
- Screening: The identified papers were screened by checking the title and abstract to eliminate those irrelevant to the field of education, evaluation, or automated scoring.
- Eligibility: All full papers were checked for eligibility based on set criteria in terms of a distinct methodology being used, technical suitability, and experimental description.
- Inclusion: Papers were considered for final evaluation only if they offered quantitative evaluation metrics like Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), correlation coefficients, etc.

Such a method adheres to a well-established and modern blueprint that has been used in recent reviews of AI-powered educational assessment. This methodology strives for a comprehensive and reproducible analysis of how architectures are developed, what kinds of data are used, and how evaluation is conducted in the AAES domain. Before performing the analysis, the review protocol defines clear criteria for categorizing architectures, datasets, and evaluation strategies to ensure consistency across the selected studies.

A. Research Questions

The survey is structured around a set of questions that provide a framework for systematically analyzing how Arabic Automated Essay Scoring (AAES) is developed, what types of data are used, how evaluation is conducted, and the future directions of AAES. These questions guide the selection, comparison, and synthesis of the literature:

RQ1: How do traditional machine learning, deep learning, transformer-based, and large language model approaches compare in terms of performance, reliability, and scoring behavior in Arabic Automated Essay Scoring.

RQ2: What kinds of datasets are being used in Arabic Automated Essay Scoring (AAES) research, and how do linguistic properties, annotation schemes, and domain factors affect model generalization and robustness.

RQ3: What kinds of prompting strategies, scoring systems, and preprocessing methods lead to more consistent, interpretable, and transparent grading in LLM-based evaluation systems.

RQ4: How do evaluation metrics—such as Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and correlation-based metrics—perform in aligning automated grades with human judgments.

RQ5: What kinds of trends in architecture, deployment, and ethics have emerged in LLM-based evaluation systems in Arabic Automated Essay Scoring.

Relevant studies were identified through a systematic search across major academic repositories, including IEEE Xplore, SpringerLink, ScienceDirect, arXiv, as well as peer-reviewed journals on educational artificial intelligence. The search scope was limited to publications between 2018 and 2026, capturing the rapid emergence of transformer architectures and LLM-based grading frameworks in this period.

A comprehensive keyword strategy was employed to ensure wide coverage of Arabic Automated Essay Scoring (AAES) research and related multilingual grading studies. The primary search terms included:

- “Arabic automated essay scoring”
- “large language models grading”
- “LLM evaluation education”
- “transformer essay scoring”
- “rubric-based prompt engineering”
- “multilingual automated grading”
- “QWK automated assessment”

These keywords were applied individually and in combination using Boolean operators to ensure comprehensive retrieval of relevant studies. The search strategy intentionally incorporated both technical, model-oriented terms and education-focused evaluation terminology to capture interdisciplinary contributions spanning natural language processing, educational measurement, and AI-assisted assessment.

A broad inclusion strategy was adopted to encompass experimental studies, methodological frameworks, and conceptual analyses addressing LLM-driven grading pipelines, multilingual evaluation frameworks, and human–AI collaboration in assessment [7], [8]. This approach ensured comprehensive coverage of emerging research trends while remaining aligned with the survey’s research questions and analytical objectives.

An example search string used in IEEE Xplore was:

("Arabic" AND "automated essay scoring") OR ("large language model" AND "educational assessment").

B. Exclusion Criteria (Any of the Following Conditions):

Studies were excluded if they met one or more of the following criteria:

- The study relied on heuristic or rule-based grading approaches without involving machine learning or neural modeling techniques.
- The study lacked experimental validation and/or did not report performance measures.
- The study focused on general natural language processing tasks unrelated to educational assessment and automated grading.

Survey and review articles were included only when they provided essential background information, theoretical underpinnings, and relevant knowledge.

C. Dataset Mapping and Organization

Each included study was carefully reviewed to identify the datasets that were employed in the construction, training, and testing of these models. Special focus was given to Arabic essay datasets such as AR-AES and other emerging Multilingual Automated Essay Scoring benchmarks.

The following aspects were examined to map these datasets:

- Language and linguistic complexity, including the intricacies of the Arabic alphabet and the distinctive morphology of the language.
- Dataset scale and scoring framework, distinguishing between analytic and holistic scoring approaches.
- Domain coverage and educational level, including essays, exams, and other structured writing tasks.
- Annotation quality and inter-rater agreement, measuring the degree of consistency between different raters in evaluating essays.

This dataset-wide mapping enables a structured comparison of architectural performance across different language models, scoring models, and annotation strategies, thus directly addressing Research Question 2. This survey also highlights significant gaps in current Arabic Automated Essay Scoring (AAES) studies, including limited dataset access, domain bias, and scoring rubric inconsistencies, all of which affect model generalization.

D. Architectural Categorization

To address Research Questions 1 and 5 (RQ1 and RQ5), the selected studies were systematically categorized by architectural paradigm. The proposed taxonomy captures the evolutionary trend of Automated Essay Scoring (AES) systems, from traditional feature-engineering models to reasoning-oriented LLM-based evaluators. Four architectural categories have been proposed for the categorization of the studies:

- Traditional Machine Learning AES: Feature-engineering models, including support vector machines, regression models, and traditional large-scale supervised learning models using handcrafted linguistic features and statistical text representations.
- Deep Learning AES: Neural networks combining convolutional and recurrent elements—such as CNN encoders, BiLSTM models, and attention mechanisms—to capture contextual information in essays.
- Transformer-Based Models: Pre-trained transformer models, such as BERT, RoBERTa, and multilingual variants, fine-tuned or task-tuned for essay scoring.
- Large Language Model Evaluators: Prompt-based LLM evaluators using instruction-tuned or generative LLMs for reasoning, interpretation, and evaluation according to educational guidelines.

The proposed architectural categorization of the models provides a unified framework for the analysis of models with different representational power, interpretability, or computational complexity. The evolutionary trend from prediction-based essay scoring toward reasoning-oriented essay evaluation models has been a dominant trend in Arabic AES system research.

Apart from the typical performance-related statistics, this survey is primarily concerned with the more specialized metrics used in Automated Essay Scoring (AES) technology. The aim is to assess the similarity between the results produced by the automated scoring system and those produced by human graders, not merely the proportion of exact matches. The key factors considered for the purpose of this assessment are:

- Quadratic Weighted Kappa (QWK): This compares the results produced by the machine scoring system with the results produced by the human scoring system, taking into consideration the distance or grading gap between the results on the ordinal scale. Because it is more sensitive to the grading gap, this is one of the more popular methods for evaluating the reliability of the scoring system.
- Mean Absolute Error (MAE): This is the average absolute difference between the predicted result and the actual result, providing insight into the accuracy or inaccuracy of the scoring system.
- Correlation Coefficients: Pearson and Spearman correlations are commonly used to evaluate the strength of alignment between automated grading outputs and human judgment patterns, offering complementary insight beyond agreement-based metrics.

Beyond predictive performance, this review also incorporates deployment-oriented evaluation factors, including computational cost, model complexity, prompt stability, and score reproducibility across multiple inference runs. These considerations are particularly relevant for large language model–based grading systems, where stochastic generation processes and high computational requirements may affect

reliability, fairness, and scalability in real-world educational settings.

E. Comparative Analysis

The final step in the review process involved the application of structured synthesis among all the studies selected in order to methodically evaluate architectural performance, dataset impacts, and grading systems. Rather than relying only on the accuracy of the grading systems, the analysis focused on methodological rigor and practical deployment considerations in Arabic Automated Essay Scoring (AAES).

Special emphasis is given to some key areas that affect the reliability of the grading system in using Large Language Model-based grading systems:

- Score stability under stochastic LLM-based inference: It examines how probabilistic generation affects reproducibility across repeated grading runs.
- Effectiveness of rubric-aligned prompting: It assesses how structured evaluation instructions improve agreement with human graders.
- Impact of multi-run scoring and statistical aggregation: It includes averaging and variance-reduction techniques designed to mitigate output variability.
- Trade-offs between interpretability and performance: It is used especially when comparing traditional machine learning models with transformer-based and generative LLM evaluators.

Model capabilities were therefore analyzed not only in terms of predictive performance but also with respect to fairness, transparency, computational feasibility, and deployment practicality in real educational environments. This multidimensional comparison enabled a more nuanced understanding of how architectural design and evaluation methodology jointly influence grading reliability.

Out of 60 studies identified in the initial phase for the candidates, 45 studies proceeded to the title and abstract screening phase after eliminating duplicates and irrelevant literature. Upon checking the full texts, 15 studies were excluded because they lacked quantitative evaluation metrics. Finally, 30 studies were included in the synthesis phase, all of which reported quantitative performance metrics related to automated essay scoring, multilingual AES, or LLM-based educational evaluation. The relatively low number of studies included is due to the fact that Arabic Automatic Essay Scoring is a fairly recent topic of investigation. Most articles excluded from the review failed to provide quantitative evaluation measures and detailed methods.

Our overall framework for synthesizing existing trends and identifying emerging trends in AI-based educational evaluation encompasses literature screening, architectural categorization, dataset analysis, and benchmark evaluation. This framework allows for a transparent and reproducible evaluation of existing and emerging trends in AI-based educational evaluation. Fig. 1 depicts the PRISMA-based flow diagram for identifying existing studies in this survey.

III. Results

In this section, the findings from the literature review are synthesized to provide a cohesive overview of how architectural factors, dataset properties, prompting strategies, and evaluation techniques combine in LLM-based Arabic Automated Essay Scoring (AAES). Rather than simply presenting discrete performance metrics, the findings are organized around the research questions to highlight the development of patterns, the effectiveness of the methods, and the challenges facing reliable grading. The text emphasizes the interplay between architectural, dataset, and prompting factors and evaluation reliability, semantic interpretation, and conformity to human evaluation.

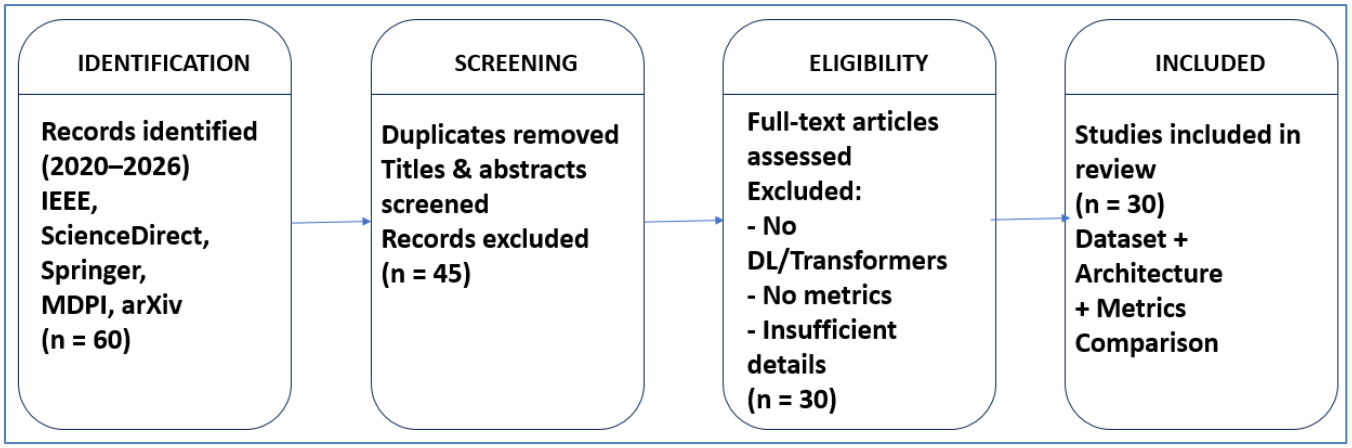


Fig. 1. PRISMA flow diagram of the study selection and analysis process.

A. Performance of Automated Essay Scoring Architectures (RQ1)

Across the surveyed studies, conventional machine learning-based AES systems tend to correlate with human evaluators, albeit at a moderate level, with correlation coefficients ranging generally from 0.55 to 0.70 [3], [4]. These approaches are also popular due to their low computational cost and ease of interpretation, but the difficulty in capturing discourse-level semantic meaning, especially in languages with complex morphology like Arabic, makes it hard to interpret the

significance of linguistic features. Then came the advent of deep learning, especially the BiLSTM and attention-based models, which improve contextual interpretation through learning hierarchical representations. These models tend to achieve improved Quadratic Weighted Kappa (QWK) scores compared to traditional machine learning approaches, reflecting better sensitivity to coherence, argument structures, and lexical variation within essays [5], [6]. Thus, the introduction of deep learning methods into the AES task marked a significant improvement over conventional machine learning methods, which relied on the engineering of linguistic features.

The introduction of Transformer-based models also took the AES task to the next level, especially with the introduction of bidirectional contextual embeddings, which capture long-range dependencies within essays. Indeed, several studies report higher QWK performance compared to earlier approaches, highlighting the effectiveness of self-attention mechanisms in capturing semantic relationships within essays [6], [9].

Recently, the advent of large language model-driven grading pipelines points to the emergence of reasoning-oriented evaluators rather than purely scoring-oriented ones. For instance, prompt-based frameworks facilitate the interpretation of grading rubrics, the generation of structured evaluations, and the provision of feedback according to the objectives.

The literature indicates that it is possible to train evaluators, such as large language models, to be as effective as human evaluators [7], [10]. However, a challenge remains: inference variability, where different evaluations can produce different results [10]. Statistical methods are needed to ensure reproducibility. Table I provides a summary of the differences, trends, and limitations of the AES architectures, including the shift from feature-oriented to reasoning-oriented evaluators, such as the LLM.

B. Dataset Characteristics and Their Influence on Model Behavior (RQ2)

Dataset characteristics play a significant role in the performance and generalization of Arabic Automated Essay Scoring (AAES) systems. In the reviewed works, factors such as dataset size, linguistic diversity, annotation practices, and scoring mechanisms are identified as the primary dataset characteristics that affect model performance and generalization.

The AR-AES dataset [9] is part of the rare set of structured datasets designed specifically for Arabic essay scoring, which supports both analytic and holistic essay scoring mechanisms. Its relatively controlled linguistic environment is advantageous for training and experimentation. This limited diversity of subjects and writing styles may affect the generalization of the AAES system, especially when used outside the original context or with different learners. On the other hand, multilingual corpora like ASAP-AES [11] are commonly employed for Automated Essay Scoring since they are large in size and include a scoring framework. However, because these corpora focus mainly on English essays, they are not directly applicable to Arabic AES due to the unique linguistic and rhetorical features. New datasets, such as ZaQQ [12], show potential for addressing shortcomings of existing datasets by combining human and AI annotation. Although these datasets aim to improve evaluation by including diverse linguistic perspectives, they are still in early development. As a result, many LLM-based grading systems are sensitive to data composition and may fail outside their original training domain on which they were initially trained.

One of the commonalities in the reviewed works is the importance of dataset diversity and annotation standards in LLM-based grading systems. The studies have shown how these systems perform poorly when faced with different writing styles, educational backgrounds, and task formats, especially when the dataset used to train the LLM is narrow in domain. Thus, increasing the size and standardization of Arabic essay datasets is essential to improve evaluation.

Table II provides a comparative analysis of the major datasets used in the research on Arabic AES, including their linguistic coverage, applications, and limitations.

C. Impact of Prompting Strategies and Scoring Frameworks (RQ3)

Prompt engineering has emerged as a key factor influencing the reliability, transparency, and reproducibility of LLM-based Automated Essay Scoring systems. Across the reviewed studies, rubric-aligned prompting consistently improved grading reliability by guiding models to evaluate essays using explicit pedagogical criteria rather than relying solely on implicit semantic similarity [7]. Structured prompts that explicitly reference scoring rubrics were shown to enhance alignment with human evaluators and reduce ambiguity in generated assessments.

Recent research also highlights the effectiveness of bilingual prompting strategies, in which Arabic evaluation instructions are complemented by English rubric descriptors. This approach leverages the strong multilingual capabilities of contemporary LLMs, improving interpretability and supporting more stable alignment with human grading standards, especially in multilingual educational contexts [8].

A notable methodological advancement is the adoption of multi-run scoring frameworks, which address the stochastic nature of generative model inference [10]. By producing multiple grading outputs and aggregating them using statistical techniques—such as averaging, variance reduction, or consensus scoring—researchers mitigate output instability and improve agreement metrics, including Quadratic Weighted Kappa (QWK) and Mean Absolute Error (MAE). These findings suggest that grading reliability is influenced not only by model architecture but also by the design of the prompting and aggregation pipeline.

Additional prompting strategies reported in the literature include:

- Few-shot prompting, where curated examples illustrate grading expectations and provide contextual guidance.
- Structured reasoning prompts, encouraging step-by-step evaluation aligned with educational rubrics and promoting explainability.

- Error-detection prompts, designed to identify inconsistencies or logical conflicts in generated scoring outputs.

Collectively, these techniques demonstrate that prompt design functions as a methodological layer comparable in importance to model architecture, directly affecting grading stability, interpretability, and fairness. Explainability is particularly important in educational assessment because grading decisions may directly influence student progression, feedback quality, and instructor trust in automated evaluation systems. Table III summarizes the primary prompting and preprocessing strategies employed in LLM-based AES systems and highlights their reported contributions to scoring reliability.

D. Evaluation Metrics and Alignment with Human Graders (RQ4)

Evaluation in Arabic Automated Essay Scoring (AAES) goes beyond conventional accuracy-based metrics to capture the nuanced relationship between automated predictions and human judgment. Across the reviewed studies, Quadratic Weighted Kappa (QWK) remains the most widely adopted evaluation measure, as it accounts for ordinal score differences and quantifies agreement between automated grading outputs and human annotations [6]. Complementary metrics, such as Mean Absolute Error (MAE), provide an interpretable measure of scoring deviation, while correlation coefficients—including Pearson and Spearman—assess linear alignment between model predictions and human grading.

LLM-based evaluators demonstrate promising performance across these evaluation criteria, particularly when integrated with rubric-driven prompting strategies and multi-run aggregation frameworks. These methodological enhancements improve agreement scores and reduce prediction variance. However, the literature emphasizes that strong numerical performance does not necessarily guarantee fairness, interpretability, or pedagogical validity. Several studies report cases in which models achieve high correlation or QWK values while producing inconsistent reasoning processes or exhibiting bias toward specific linguistic styles, essay lengths, or structural patterns [7], [8].

As a result, recent research increasingly advocates for complementary evaluation paradigms that incorporate qualitative analysis of model-generated feedback, human-centered evaluation protocols, and transparency-oriented assessment frameworks. These approaches aim to bridge the gap between quantitative agreement metrics and broader educational objectives of automated assessment systems.

A comparative overview of computational and evaluation trade-offs across AES architectures is presented in Table IV, highlighting how differences in model complexity, prompting strategies, and inference mechanisms influence both grading performance and practical deployment considerations.

In order to provide a quantitative analysis of architectural performance trends in the reviewed AES systems, Table V presents an overview of the evaluation measures, performance metrics, and trade-offs encountered by each of the leading architectures in AES.

According to Table V, AES based on transformers and large language models demonstrates superior agreement with human graders compared to machine-learning-based AES systems. Nevertheless, they are characterized by higher computational complexity and instability of grading.

E. Emerging Trends and Opportunities in Arabic AES (RQ5)

The literature review revealed several emerging trends that are changing how Arabic automated essay scoring with AI technology is carried out. Overall, there is a shift from traditional prediction-based grading systems to approaches that focus more on reasoning and human-centered evaluation.

The first trend is the shift towards reasoning-based grading systems that use large language models. Compared to traditional feature-based or semantic grading systems, large language models better understand context and provide more accurate evaluations for essays in different languages [7], [8].

The second trend is the focus on human-AI collaborative grading systems. Compared to AI-only grading systems, recent approaches focus on human-AI collaborative systems where large language models assist in grading essays and provide explanations for their decisions [10].

The third trend is the focus on explainability, fairness, and responsible use in Arabic automated essay grading systems. Compared to AI-only grading systems for Arabic essays, recent approaches emphasize prompt transparency and bias detection in large language models.

The fourth trend is the focus on developing multilingual foundation models for Arabic automated essay grading systems. Compared to LLM-only grading systems for Arabic essays, recent approaches aim to develop more robust systems capable of adapting to different educational contexts and environments [9].

The assembled evidence suggests that there is a paradigm shift in the field of automated essay scoring, from machine learning pipelines based on feature engineering to reasoning-oriented evaluation systems facilitated by the use of transformer models and large language models. This shift is part of the larger paradigm transition in the field of educational AI, from predictive scoring models to more semantically grounded models of evaluation that facilitate reasoning and the production of structured feedback based on rubrics. While the transformer and LLM models have shown clear improvements in contextual understanding and human-aligned evaluation, they have also led to the identification of methodological and deployment issues related to reproducibility in the presence of stochasticity, computational demands, and the identification of issues related to the ethics of deployment as core concerns in the deployment of Arabic Automated Essay Scoring systems.

F. Balancing Linguistic Understanding and Grading Reliability

A notable feature of the studies examined here is the intrinsic trade-off between the semantic reasoning capacity and the scoring stability of the automated essay scoring systems. Conventional machine learning techniques, such as support vector machines and regression-based grading models, offer good interpretability and efficiency; however, the feature

engineering limitations of such systems hinder the accurate modeling of discourse semantics and contextual relationships, especially for morphologically rich languages such as Arabic [3], [13]. Hence, while such systems frequently demonstrate scoring stability, the outputs may not generalize well for assessing higher-order writing competencies such as argumentation quality, coherence, and semantics.

The adoption of deep learning and transformer-based models has greatly improved the contextual modeling capacity of the essay scoring systems. Pre-trained models like AraBERT and multilingual transformers improve the model's capacity to grasp the meaning of words, which improves the alignment of the automated scoring system with the scoring of human graders [6], [9]. However, it is essential to note that the models rely on the quality of the datasets. This is because the datasets for the Arabic education domain are limited. This implies that the model's performance in an experimental setting does not necessarily reflect the real-world situation.

Large language models (LLMs) offer a further advancement for automated essay scoring by supporting reasoning-based evaluation rather than purely predictive scoring. When guided by well-designed prompts, LLMs can interpret scoring criteria, generate structured explanations, and provide pedagogically meaningful feedback [7], [10]. Nevertheless, because these models operate probabilistically, repeated evaluations may produce variable results, raising concerns about consistency and reproducibility in grading.

TABLE I. COMPARISON OF AUTOMATED ESSAY SCORING ARCHITECTURES

Ref	Model category	Example architectures	Performance trend	Primary application	Limitations
[2], [14]	Traditional machine learning	SVM, regression-based AES	Moderate agreement with human graders	Educational analytics	Limited contextual understanding
[3], [4]	Deep learning models	BiLSTM, CNN–LSTM	Improved contextual evaluation	Structured essay grading	Requires large labeled datasets
[6], [9]	Transformer-based models	BERT, AraBERT, multilingual transformers	Strong semantic representation	Multilingual and Arabic AES	Computationally expensive
[15]	Hybrid architectures	Multi-model fusion, ensemble learning	Stable grading performance	Cross-dataset AES evaluation	Complex training pipelines
[16], [17]	Large language models	GPT-based and instruction-tuned LLMs	High reasoning and alignment potential	Rubric-guided scoring	Output variability

TABLE II. DATASETS USED IN ARABIC AND CROSS-LINGUAL AES RESEARCH

Ref	Dataset	Language	Task type	Typical usage	Identified issues
[12]	ZaQQ dataset	Arabic	Human–AI collaborative AES	LLM evaluation research	Limited scale
[13], [18]	ASAP-AES benchmark	English	Holistic essay scoring	Model benchmarking	Not Arabic-native
[19]	Arabic essay corpus	Arabic	Grammar-based scoring	Educational assessment	Annotation inconsistency
[9]	Arabic grammatical AES	Arabic	Linguistic evaluation	Transformer-based grading	Domain-specific bias
[20]	Multilingual AES datasets	Mixed languages	Cross-lingual grading	LLM evaluation studies	Domain shift

TABLE III. PROMPTING AND PREPROCESSING STRATEGIES

Ref	Technique	Reported improvement	Affected models	Observations
[7], [21]	Rubric-guided prompting	Improved human alignment	GPT-style graders	Enhances grading consistency
[10]	Few-shot prompting	Reduced scoring variance	Instruction-tuned LLMs	Provides contextual guidance
[22], [23]	Multi-run evaluation	Stabilized grading outputs	LLM evaluators	Mitigates stochastic behavior
[9]	Arabic text normalization	Better semantic accuracy	Transformer-based AES	Handles morphology and dialects
[24], [25]	Structured reasoning prompts	Improved transparency	Generative AI graders	Supports ethical AI evaluation

TABLE IV. COMPUTATIONAL TRADE-OFFS ACROSS AES ARCHITECTURES

Ref	Model	Approximate parameters	Computational cost	Evaluation metrics	Deployment suitability
[3], [13]	SVM-based AES	Low	Low CPU usage	MAE, correlation	High for schools
[6]	Transformer AES (BERT)	~110M	Moderate GPU usage	QWK, F1-score	Medium
[9]	Arabic transformer AES	~120M	Moderate	QWK, accuracy	Medium
[7], [26]	GPT-style LLM grading	Billions	Very high cloud cost	Human alignment metrics	Low for edge deployment
[20], [27]	Distilled LLM AES	Reduced parameters	Medium	QWK, MAE	High future potential

TABLE V. QUANTITATIVE PERFORMANCE TRENDS ACROSS AES ARCHITECTURES

Ref	Architecture	Typical Metrics	Performance Trend	Main Limitation
[3], [4]	Traditional ML	MAE, Correlation	Moderate agreement	Weak semantic understanding
[5], [6]	Deep Learning	QWK, MAE	Improved contextual scoring	Requires large datasets
[6], [9]	Transformer-based	QWK, Accuracy	Strong semantic alignment	High computation
[7], [10]	LLM-based AES	QWK, Human alignment	Strong reasoning capability	Output variability

G. Dataset Limitations and Generalization Challenges

One of the difficulties associated with AAES studies involves the scarcity of Arabic essay corpora. Although corpus-based models of English essays, such as the ASAP-AES dataset, help compare the performance of various grading algorithms in AAES, the nature of the texts in those corpora differs greatly from Arabic writing. AAES, too, faces many NLP challenges associated with the Arabic language, such as morphological richness, spelling/orthography variation, the presence or absence of diacritics, and regional dialectal variations. This presents a problem for transformer-based and LLM-based AAES, which are typically built using only MSA-based and multilingual corpora. Moreover, the current Arabic educational corpora used in AAES research are not sufficiently large or diverse [6], [9].

As the literature surveyed in the previous sections shows, the way a system behaves is heavily influenced by the evaluation resource it was exposed to. Inconsistencies in the way the resources are labeled may cause difficulties in reproducing results from one paper to another. Domain shifts, such as the one between essay prompts and actual student responses to these prompts, also cause difficulties in generalizing the results. In these cases, the performance tends to degrade as the system is exposed to new contexts.

Nevertheless, the recent advances in large language models are alleviating some of the difficulties researchers face in the field. Large language models generalize well to other semantic spaces and enhance the performance of the system on cross-lingual and cross-genre tasks [8]. However, the lack of good resources for the Arabic essay domain still poses a significant barrier to the evaluation process.

H. Influence of Prompting Strategies and Preprocessing

Prompt design is currently at the center of the effectiveness and reliability of automated essay scoring models that use large language models. Current research shows that prompt-based methods that align with a rubric and use thoughtfully designed few-shot examples allow models to better align their scores with human raters since the model is designed to evaluate the student response according to a specific framework. By integrating the scoring rubric into the prompt, these prompt-based methods force the model to evaluate the student response more thoughtfully rather than simply relying on language generation. When combined with other prompt-based methods that incorporate reasoning-based prompts, the process is more transparent since the model will outline the steps it takes to evaluate the student response. This is in line with new AI ethics that focus on accountability and transparency [25].

However, it is also worth noting that preprocessing methods that take into account the nuances of the Arabic language also greatly improve the accuracy of the semantic understanding and grading. By taking into account issues such as orthographic normalization and morphology-based tokenization, the model is better able to filter out irrelevant information that arises from variations in spelling and morphology. This is particularly important for models that use a transformer-based architecture and a large language model since token-level information is particularly important for the final grading. However, prompt-based methods are still a source of variability for LLM-based AES models. The fact that small changes to the prompt and scoring rubric can result in a change to the final score also brings into question the reproducibility and fairness of the grading process. Recent studies also highlight risks related to hallucinated feedback, prompt sensitivity, and bias amplification, where slight prompt variations or linguistic patterns may influence grading consistency and fairness across student populations.

I. Computational Cost and Deployment Feasibility

This is particularly the case despite the marked improvements in terms of performance that have been reported in the literature on the use of LLMs in AES systems. For example, the use of external inference services is not transparent in terms of the exact ways in which the models function, and it also limits the extent of control that can be exercised by the educational institutions in question, which is particularly problematic in terms of governance and other issues of relevance.

From the point of view of computational complexity, classical machine learning models in the AES system remain attractive, despite their limitations in terms of their ability to reason and their low computational requirements [3]. On the other hand, the Transformer model offers contextual evaluation efficiently, without the computational costs that have been reported in the application of LLMs in the AES system.

Despite the clear improvements in the application of LLMs in the AES system, the literature indicates that there are high computational requirements, even when cloud computing and/or hardware accelerators are used [7], [26]. Moreover, the challenges associated with the use of external APIs have led to the development of techniques for efficient fine-tuning, reducing the computational requirements without compromising the performance in terms of the ability to reason.

The literature indicates the need to balance the expressiveness of the model and its practicality, especially in the application of LLMs in the AES system, in terms of the ability to reason and the computational requirements.

J. Emerging Trends in AI-Based Educational Assessment

The results of this survey indicate that there are several trends that are currently shaping the way in which educational assessments with AI will be carried out in the future. One of the trends that has been identified in this regard is that of hybrid assessments that use a combination of large language model-based reasoning with other scoring mechanisms. Instead of purely relying on the output of the language model, hybrid assessments use rule-based metrics, statistical features, and rubric-based prompts to make grading easier to understand. By using this approach, hybrid assessments have been shown to eliminate randomness in grading without losing the flexibility of language model-based grading.

Another trend that has been identified in this regard is that of multilingual and multi-cultural assessments. As educational assessments with AI move beyond purely English-based datasets, there is an emerging need to create grading models that take into account linguistic diversity, dialectical diversity, and educational diversity. Multilingual assessments focus on adapting grading criteria to specific educational contexts within countries while maintaining reliability across languages [11]. This trend is in keeping with the need to create globally accessible AI-based assessments.

Ethics-based governance is also an emerging trend in this regard. The Artificial Intelligence Assessment Scale has been developed to create guidelines that encourage explainability, bias detection, and responsible AI-based education [24], [25].

Recent discussions also emphasize fairness across Arabic dialects, mitigation of cultural bias in grading, and protection of student educational data privacy during large-scale AI-assisted assessment. This is particularly important in grading contexts since grading decisions often determine student progression. As such, new research in this field has shown that there is an emerging trend of human-centered assessments that use explainability, auditability, and pedagogical oversight in language model-based grading pipelines.

K. Implications for Future Arabic AES Research

The survey indicates several areas that should be pursued to further advance Arabic Automated Essay Scoring, so that it can eventually work well in classrooms. First off, there is a need to develop larger and more diverse Arabic essay datasets. Current datasets are small, not diverse in topics, and the annotation quality is inconsistent. Future datasets must include standardized scoring rubrics, multi-rater annotation, and cover diverse settings to allow reproducible benchmarking.

Another important research direction involves robust prompting methods and evaluation benchmarks to stabilize Arabic AES with large language models. Prompts are known to significantly affect generative model outputs, so standardized rubric-aligned prompting frameworks should be developed, methods to aggregate results over several runs, and transparent reporting. Developing shared evaluation methods will allow us to better compare architectures, which will help unify the field.

The third area that should be pursued is multimodal inputs in Arabic AES. Multimodal inputs, which include other signals beyond text, such as handwriting, student interaction, or curriculum information, could enable Arabic AES to assess more than text quality to better understand other factors. This is in line with other areas of educational AI, which are moving toward more holistic evaluation methods, not just text-based evaluation.

The last area that should be pursued is addressing practical challenges, beyond peak performance metrics. While most studies focus on higher QWK or correlation scores, few studies investigate other important factors, such as computational cost, data privacy, scalability of infrastructure, and usability of Arabic AES. These factors are important to transition Arabic AES from experimental settings to practical and deployable educational technology.

For future research on Arabic AES, retrieval-augmented grading framework, specialized Arabic LLMs in addition to parameter-efficient fine-tuning methods tailored to educational purposes can be adopted. Moreover, collaborative evaluation models that involve humans in the loop may offer better clarity in grading while still enabling human control over the process.

L. Limitations of the Survey

This review aggregates the current progress in Arabic Automated Essay Scoring (AAES) systems, but there are several caveats that should be noted. The process of selecting the relevant literature heavily relied on indexed digital libraries, which may introduce publication bias or omit newer regional literature. The findings in the literature have been based on reported outcomes and not a unified benchmark, which makes it difficult to compare the literature due to differences in the datasets used, the rubrics for scoring, and the methods of evaluation. The field of large language models is also evolving at a rapid rate, which further limits the temporal relevance of the literature, with newer models and methods being developed frequently. The lack of large standardized datasets of essays in Arabic is also a problem that limits the generalization of the literature.

IV. Conclusion

This review distills the latest advances in Arabic Automated Essay Scoring to demonstrate how new developments in transformer-based models and large language models are revolutionizing the field of automated education assessment. The results of the research demonstrate that recent advances in natural language processing have revolutionized the field of automated grading to an unprecedented level of fidelity in determining semantic coherence, language quality, and grading rubric alignment. While the older approach to automated grading relied on the use of various features to arrive at an assessment of the student response to an assignment, the new approach of using transformer-based models and large language models has clearly ushered in a new era of grading that is more context-oriented and reasoning-based.

However, the results of the research also demonstrate that while there has been an unprecedented level of progress in the field of AAES, there remains much to be addressed in terms of the complexity of the task. One of the problems is that there is currently a bottleneck in terms of the amount of data available in the Arabic language to support AAES. While domain variety in available datasets is limited to support AAES, there is also an inconsistency in the annotations of the available data. While

multilingual pretraining can be used to overcome some of these problems to an extent, language diversity remains a critical factor in determining the robustness of grading.

The survey also shines a light on how prompting strategies and preprocessing pipelines are at the core of how effective an evaluation carried out by an LLM-based evaluator turns out to be. While strategies such as rubric-aware prompting, structured reasoning prompts, and Arabic normalization have proven to be effective in improving grading clarity and consistency, however, these generative models remain highly sensitive to prompting strategies. This is an area that calls for standardized prompting strategies to be developed. However, there is no doubt that ethical considerations have now become an important part of how AI-based evaluators contribute to the educational field. This is because of the broader developments in the direction of developing trustworthy AI in education.

From an implementational perspective, a balance must be maintained between evaluator effectiveness and deployment practicality of the model in an educational setting. While traditional ML approaches and lightweight transformer-based models can be attractive for large-scale educational settings because of their practicality, ease of interpretability, and predictability, there is also the possibility that an LLM-based evaluator can be more effective in its reasoning and feedback mechanisms. However, there is also the flip side of how computationally intensive they can be to deploy, with potential variability in scores due to model stochasticity. Recent developments in parameter-efficient fine-tuning and distillation-based approaches suggest promising avenues to achieve high performance while maintaining deployability.

Looking forward into the future, it is possible to highlight a few research areas that are considered crucial for advancing Arabic automated essay scoring. For example, creating large standardized datasets for essays written in Arabic with standardized annotation schemes will be instrumental in advancing robustness and enabling comparative evaluations across different models. Another direction that may minimize random variability in evaluation outcomes while maintaining interpretability is the use of hybrid grading models that combine symbolic evaluation with generative reasoning. Finally, integrating multimodal educational signals, including handwriting recognition, student metadata, and broader learning analytics, may be very beneficial in advancing evaluation accuracy and personalization.

To conclude, the shift toward LLM-based evaluation for evaluation in place of more traditional automated evaluation systems represents a significant leap forward in AI-based evaluation in educational settings. The current approaches already point toward significant scalability and support for teachers in evaluation processes. However, future research should emphasize robustness, fairness, and interpretability in advancing Arabic automated essay scoring in order to ensure that it moves beyond experimental systems toward more robust systems that align with pedagogical needs and support modernization in evaluation processes.

REFERENCES

[1] H. Niemi, R. D. Pea, and Y. Lu, "Introduction to AI in Learning: Designing the Future," in AI in Learning: Designing the Future, Springer, 2022, pp. 1–15. doi: 10.1007/978-3-031-09687-7_1.

[2] D. S. Yadav, "Navigating the Landscape of AI Integration in Education: Opportunities, Challenges, and Ethical Considerations for Harnessing the Potential of Artificial Intelligence for Teaching and Learning," BSSS Journal of Computer, vol. 15, no. 1, pp. 38–48, 2024, doi: 10.51767/jc1503.

[3] D. S. V. Madala, A. Gangal, S. Krishna, A. Goyal, and A. Sureka, "An Empirical Analysis of Machine Learning Models for Automated Essay Grading." 2018. doi: 10.7287/peerj.preprints.3518v1.

[4] V. V. Ramalingam, A. Pandian, P. Chetry, and H. Nigam, "Automated Essay Grading Using Machine Learning Algorithm," Journal of Physics: Conference Series, 2018, doi: 10.1088/1742-6596/1000/1/012030.

[5] G. Liang, B. W. On, D. Jeong, H. C. Kim, and G. S. Choi, "Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture," Symmetry, vol. 10, no. 12, 2018, doi: 10.3390/sym10120682.

[6] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated Essay Scoring Using Transformer Models," Psych, vol. 3, no. 4, pp. 897–915, 2021, doi: 10.3390/psych3040056.

[7] A. Gandolfi, "GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions," International Journal of Artificial Intelligence in Education, vol. 35, no. 1, pp. 367–397, 2025, doi: 10.1007/s40593-024-00403-3.

[8] I. D. Mienye, N. Jere, G. Obaido, O. O. Ogunruku, E. Esenogho, and C. Modisane, "Large Language Models: An Overview of Foundational Architectures, Recent Trends, and a New Taxonomy," Springer Nature, 2025, doi: 10.1007/s42452-025-07668-w.

[9] S. Mahmoud, E. Nabil, and M. Torki, "Automatic Scoring of Arabic Essays: A Parameter-Efficient Approach for Grammatical Assessment," IEEE Access, vol. 12, pp. 142555–142568, 2024, doi: 10.1109/ACCESS.2024.3470728.

[10] A. Kundu and D. Barbosa, "Are Large Language Models Good Essay Graders?" Sep. 2024. [Online]. Available: http://arxiv.org/abs/2409.13120

[11] Z. Ke and V. Ng, "Automated Essay Scoring: A Survey of the State of the Art," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6300–6308. doi: 10.24963/ijcai.2019/879.

[12] Y. Elsayed, E. Nabil, M. Torki, S. Faizullah, and A. Khalafallah, "ZaQQ: A New Arabic Dataset for Automatic Essay Scoring via a Novel Human–AI Collaborative Framework," Data, vol. 10, no. 9, 2025, doi: 10.3390/data10090148.

[13] A. Singh, "Automated Essay Scoring Using Machine Learning," International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT), 2019, [Online]. Available: https://www.ijariit.com

[14] T. Zhao, "AI in Educational Technology." 2023. doi: 10.20944/preprints202311.0106.v1.

[15] M. Uto, I. Aomi, E. Tsutsumi, and M. Ueno, "Integration of Prediction Scores from Various Automated Essay Scoring Models Using Item Response Theory," IEEE Transactions on Learning Technologies, vol. 16, no. 6, pp. 983–1000, 2023, doi: 10.1109/TLT.2023.3253215.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Computing Surveys, vol. 55, no. 9, pp. 1–35, 2023, doi: 10.1145/3560815.

[17] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023, [Online]. Available: https://arxiv.org/abs/2303.08774

[18] H. Li, C. H. Chen, K. Fan, C. Young-Johnson, S. Lim, and Y. Feng, "Agreement Between Large Language Models and Human Raters in Essay Scoring: A Research Synthesis," arXiv preprint arXiv:2512.14561, 2025, [Online]. Available: https://arxiv.org/abs/2512.14561

[19] S. Esmaile, O. Al-Awaida, and B. Alshargabi, "Automated Arabic Essay Grading System based on Support Vector Machine and Text Similarity Algorithm," International Journal of Advanced Computer Science and Applications, 2019, [Online]. Available: https://meu.edu.jo/libraryTheses/5d3c0808d27b7_1.pdf

[20] R. F. Mello et al., "Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?," in Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25), ACM, 2025, pp. 93–103. doi: 10.1145/3706468.3706481.

[21] C. Grévisse, "LLM-based automatic short answer grading in undergraduate medical education," BMC Medical Education, vol. 24, no. 1, p. 1060, 2024, doi: 10.1186/s12909-024-06026-5.

[22] R. Ghazawi and E. Simpson, "Automated Essay Scoring in Arabic: A Dataset and Analysis of a BERT-based System," arXiv preprint arXiv:2407.11212, 2024, [Online]. Available: https://arxiv.org/abs/2407.11212

[23] W. Mansour, S. Albatarni, S. Eltanbouly, and T. Elsayed, "Can Large Language Models Automatically Score Proficiency of Written Essays?" Apr. 2024. [Online]. Available: http://arxiv.org/abs/2403.06149

[24] E. D. Lindsay, M. Zhang, A. Johri, and J. Bjerva, "The Responsible Development of Automated Student Feedback with Generative AI," in EDUCON, 2025. doi: 10.1109/EDUCON62633.2025.11016572.

[25] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, "The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment," Journal of University Teaching and Learning Practice, vol. 21, no. 6, 2024, doi: 10.53761/q3azde36.

[26] Z. I. Karsa and B. Goldschmidt, "Automatic Evaluation of Programming Tasks Supported by Language Models," IEEE Access, vol. 13, pp. 147741–147756, 2025, doi: 10.1109/ACCESS.2025.3601448.

[27] K. Ono and A. Morita, "Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU," Mar. 2024, doi: 10.36227/techrxiv.170956672.21573677/v1.

Revolutionizing Urban Waste Systems: A Comprehensive Review of IoT, AI, and Optimization Techniques for Smart Waste Management

Lahcen GOUSKIR[1]*, Abdelmoula ABOUHILAL[2], Mohamed GOUSKIR[3],
Mohamed BASLAM[4], Soufiane BELHOUIDEG[5], Hanaa Hachimi[6]
Faculty of Science and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco[1, 4]
National School of Applied Sciences, Sultan Moulay Slimane University, Beni Mellal, Morocco[3, 5]
Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco[2]
National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco[6]

***Abstract*—Urban waste management is shifting from fixed, reactive collection toward data-driven and adaptive service models. This review synthesizes 33 recent studies and technical contributions on Artificial Intelligence of Things (AIoT) for smart waste management. The synthesis is organized around six analytical dimensions: IoT sensing and communication, real-time monitoring performance, dynamic routing, AI-based classification and robotic sorting, edge/fog/cloud intelligence, and circular governance. Unlike a descriptive survey, the paper develops an AIoT-SWM taxonomy and an evaluation rubric for comparing smart waste systems according to interoperability, latency, energy profile, scalability, decision autonomy, circular-economy contribution, and social inclusion. Reported deployments show measurable operational gains, including 26-35% reductions in fuel consumption, more than 30% improvement in fleet utilization, 25% gains in collection efficiency, and up to 48% reduction in overflow incidents. AI sorting studies also report controlled classification accuracies above 90%, while recent techno-economic evidence indicates 95.1% material purity, 50 items/minute throughput, and payback periods of 4.3-4.9 years under specific emerging-economy conditions. The review concludes that AIoT can improve municipal waste services only when technical performance guarantees are combined with open data standards, cybersecurity safeguards, human oversight, and context-sensitive inclusion of informal waste actors.**

Keywords*—Smart waste management; Internet of Things; Artificial Intelligence; AIoT; edge computing; dynamic routing; circular economy; robotic sorting

I. Introduction

A. Urban Waste Pressure and the Need for Intelligent Systems

Rapid urbanization and changes in consumption have increased the volume and heterogeneity of municipal solid waste (MSW). Many cities still depend on fixed schedules, manual reporting, and weakly integrated disposal chains. These models are costly and slow because they do not observe real-time waste generation, traffic conditions, or neighborhood-level variability [2], [4].

Smart waste management is therefore not only a matter of installing sensors. It requires a cyber-physical architecture that can sense the state of containers, transmit data reliably, transform data into operational decisions, and connect waste services with recycling, public health, mobility, and environmental governance [3], [13], [18], [24], [25].

The unresolved research problem is systems-level integration. Existing deployments often perform well as isolated pilots but remain limited by vendor lock-in, poor interoperability, uncertain latency guarantees, energy constraints, and weak governance of data and algorithms [10], [19], [21]. These gaps explain why many smart-bin pilots do not scale into citywide waste intelligence.

B. Scope and Contribution of the Review

This review examines how IoT, AI, optimization, robotic sorting, blockchain, and edge computing can support a transition from linear waste handling to adaptive and circular urban waste systems. The review focuses on municipal solid waste and closely related urban waste streams, including e-waste and recyclable fractions.

The main contribution is a structured synthesis artifact rather than a new experiment. The article proposes an AIoT-SWM taxonomy, a decision-oriented evaluation rubric, and a ranked research agenda. These instruments help compare systems across technical, economic, environmental, and social dimensions. They also clarify the trade-offs that municipalities face when moving from pilot projects to scalable infrastructure.

- Taxonomy: legacy, digitized, and AIoT-enabled waste systems are compared across sensing, connectivity, analytics, autonomy, and governance maturity.
- Evaluation rubric: smart waste systems are assessed using interoperability, latency, energy consumption, density, scalability, circularity, cybersecurity, and inclusion criteria.
- Research agenda: challenges are ranked by technical severity, deployment frequency, and urgency for future research and municipal policy.

*Corresponding author.

II. Review Methodology, Related Work, and Positioning

A. Review Design and Selection Criteria

The review follows a structured narrative approach. It prioritizes peer-reviewed studies and recent technical contributions published between 2020 and 2026 in the areas of smart waste management, IoT-enabled collection, AI sorting, routing optimization, edge/fog/cloud computing, blockchain traceability, and circular economy. Studies were retained when their primary domain was municipal or urban waste management, or when their method directly supported a waste-management function such as bin monitoring, vehicle routing, sorting, material recovery, or governance.

To improve citation relevance, studies from unrelated application domains were excluded from the evidence base because they did not provide traceable evidence for urban waste management. As summarized in Table I, the revised bibliography prioritizes domain-specific references on LoRaWAN smart waste infrastructure, IoT smart bins, CNN-based smart waste classification, 5G-enabled smart waste systems, and robotic sorting feasibility [18], [22], [23], [28], [33].

TABLE I. Structured Review Criteria and Synthesis Outputs

Dimension	Inclusion Focus	Synthesis Output
Domain relevance	Municipal solid waste, smart bins, collection, sorting, recycling, e-waste, circular waste systems	Removal of non-waste references and strengthened citation base
Technology layer	Sensors, communication protocols, edge/fog/cloud, AI models, routing algorithms, robotic systems	AIoT-SWM architecture and technology taxonomy
Performance evidence	Latency, accuracy, throughput, energy, packet reliability, deployment density, fuel and route efficiency	Quantitative comparison tables and timing requirements
Decision and governance	MCDM, interoperability, cybersecurity, policy, inclusion, circular economy	Evaluation rubric and ranked research agenda
Practical feasibility	Cost, maintenance, calibration, integration complexity, human oversight	Balanced techno-economic assessment of robotic and AIoT systems

B. Related Work and Differentiation

Recent reviews and impact studies show that AI and IoT improve collection, sorting, recycling, and monitoring, but many remain broad and technology-centered [11], [29], [30]. Reviews on smart bins and circular resource recovery focus strongly on sensing and recovery infrastructure [13], whereas machine-learning reviews emphasize prediction, classification, and resource valorization [12], [15], [16].

The present review differs by connecting these strands into a systems-level framework. It does not treat IoT, AI, routing, edge computing, and circularity as separate themes. It maps their dependencies, identifies operational trade-offs, and proposes measurable criteria for deployment evaluation. As summarized in Table II, this positioning responds to the need for reviews that are useful for both researchers and municipal decision makers.

TABLE II. Positioning Against Existing Reviews

Review Stream	Dominant Focus	Gap Addressed in This Paper
AI/IoT review of collection, sorting, and recycling	General benefits of AI and IoT for waste operations [30]	Adds latency, interoperability, and deployment-readiness criteria
Smart bins for resource recovery	Smart-bin deployment and recovery outcomes [13]	Connects bin sensing to routing, edge intelligence, and circular governance
Machine learning for circular waste systems	Prediction, classification, and material recovery [12], [15]	Integrates ML with communication constraints, human oversight, and equity risks
IoT-enabled smart waste systematic reviews	Architectural components and city applications [29]	Adds a decision rubric and a ranked research agenda for scalable municipal adoption

III. Evolution of Waste Management Systems

A. From Linear Logistics to AIoT Ecosystems

Legacy MSW systems are built around collection, transport, and disposal. They offer limited visibility into fill levels, material composition, route status, or environmental risk. Their main advantage is operational simplicity, but their main weakness is that they respond after problems appear.

Digitized systems added GPS tracking, dashboards, GIS, and digital work orders. This improved transparency, but many systems remained descriptive rather than predictive. The major bottleneck was not data availability alone; it was the absence of interoperable architecture and decision intelligence [26].

AIoT-enabled systems combine sensor telemetry, edge analytics, cloud training, optimization algorithms, and decision dashboards. They can predict overflow, update routes, identify contamination, and support circular-economy reporting. As shown in Table III, their main risk is increased complexity: poor cybersecurity, weak standards, and insufficient maintenance can turn smart infrastructure into fragile infrastructure.

TABLE III. COMPARATIVE EVOLUTION OF URBAN WASTE SYSTEM ARCHITECTURES

Metric	Legacy Systems	Digitized Systems	AIoT-Enabled Systems
Main data source	Manual logs, fixed schedules, operator reports	GPS, GIS, digital dashboards	Real-time sensor telemetry, AI inference, multi-source data
Decision logic	Routine-based and reactive	Rule-based monitoring	Predictive and optimization-based
Latency profile	Hours to days	Minutes to hours	Seconds to minutes depending on protocol and edge layer
Scalability constraint	Labor, fleet, landfill capacity	Data silos and proprietary tools	Interoperability, cybersecurity, and maintenance capacity
Circular contribution	Low recovery and weak traceability	Improved reporting but limited automation	Material classification, recovery forecasting, traceability, and incentives
Main trade-off	Low complexity but low responsiveness	Visibility without autonomy	High performance but higher integration and governance burden

IV. IOT INFRASTRUCTURE IN WASTE MANAGEMENT

A. Sensor Modalities and Data Reliability

IoT infrastructure begins with the instrumentation of bins, vehicles, and facilities. Ultrasonic sensors estimate fill levels; load cells measure weight; gas sensors detect decomposition or hazardous emissions; temperature sensors support fire-risk monitoring; RFID and QR identifiers link bins, users, or assets to service records [18], [20], [22].

Sensor fusion is necessary because no single sensor is robust in all waste environments. Ultrasonic readings can be affected by irregular waste surfaces, condensation, or acoustic noise. Weight sensors may be distorted by density variation across waste types. Gas sensors drift over time and require calibration. Combining fill level, weight, temperature, and gas signals increase confidence and allows edge-side anomaly detection.

B. Communication Protocols and Deployment Trade-Offs

Communication choices determine the practical scale of smart waste systems. Long-range protocols such as LoRaWAN are suitable for low-rate telemetry across large urban areas, while NB-IoT provides cellular-grade coverage and higher device density. ZigBee and BLE are useful for short-range clusters but require gateways. Wi-Fi or 5G may support cameras and higher data rates, but their energy and cost profiles are less suitable for small battery-powered bins [18], [28], [31].

TABLE IV. PRACTICAL PROTOCOL COMPARISON FOR SMART WASTE DEPLOYMENTS

Protocol	Typical Range	Latency Profile	Energy Profile	Deployment Density	Best Waste Use Case
LoRaWAN	Urban: hundreds of meters to several km; rural/peri-urban: longer	Seconds to minutes; not ideal for continuous video	Very low for small periodic packets	High, if duty-cycle and gateway planning are respected	Fill-level telemetry, temperature alerts, low-rate bin monitoring
NB-IoT	Cellular coverage with good indoor penetration	Usually seconds; operator dependent	Low to medium	Very high device density through licensed networks	Dense urban cores, underground bins, high-rise waste rooms
ZigBee/BLE Mesh	Short range, typically building/campus scale	Milliseconds to seconds inside mesh	Low	High locally but gateway dependent	Campus bins, indoor recycling rooms, industrial parks
Wi-Fi/5G	Access-point/cellular dependent	Low latency and higher throughput	Medium to high	Good where infrastructure exists	Camera-based sorting, dashboards, fleet depots, campus pilots

As summarized in Table IV, the choice of protocol must be driven by the required message size, alert urgency, node density, and power budget. For example, a fill-level packet can be transmitted every 15-60 minutes through LoRaWAN, while a fire or hazardous-gas alarm requires priority transmission and local edge triggering. Video-based sorting should not normally stream continuously from public bins; edge inference can transmit only events or metadata to preserve bandwidth and privacy [23].

C. Power Management and Sensor Longevity

Energy management is a primary condition for citywide feasibility. Battery-powered smart bins must minimize sensing, computation, and transmission time. Duty cycling, local aggregation, event-triggered messaging, and payload compression reduce energy consumption. Solar augmentation can extend the lifetime, but is site-dependent because shading, vandalism, dust, and seasonal variation reduce effective generation.

The revised framework, therefore treats energy as a design constraint rather than a peripheral issue. A technically successful pilot cannot be considered scalable if maintenance teams must frequently replace batteries, recalibrate sensors, or repair exposed electronics.

Fig. 1 summarizes the distributed architecture used in this review. Sensors collect local state information. Edge nodes filter and classify events close to the data source. Fog gateways aggregate neighborhood-level signals and support local coordination. Cloud services support long-term storage, model training, strategic dashboards, and cross-city comparison. This distribution improves resilience because the system can maintain partial service during network degradation or cloud outages.

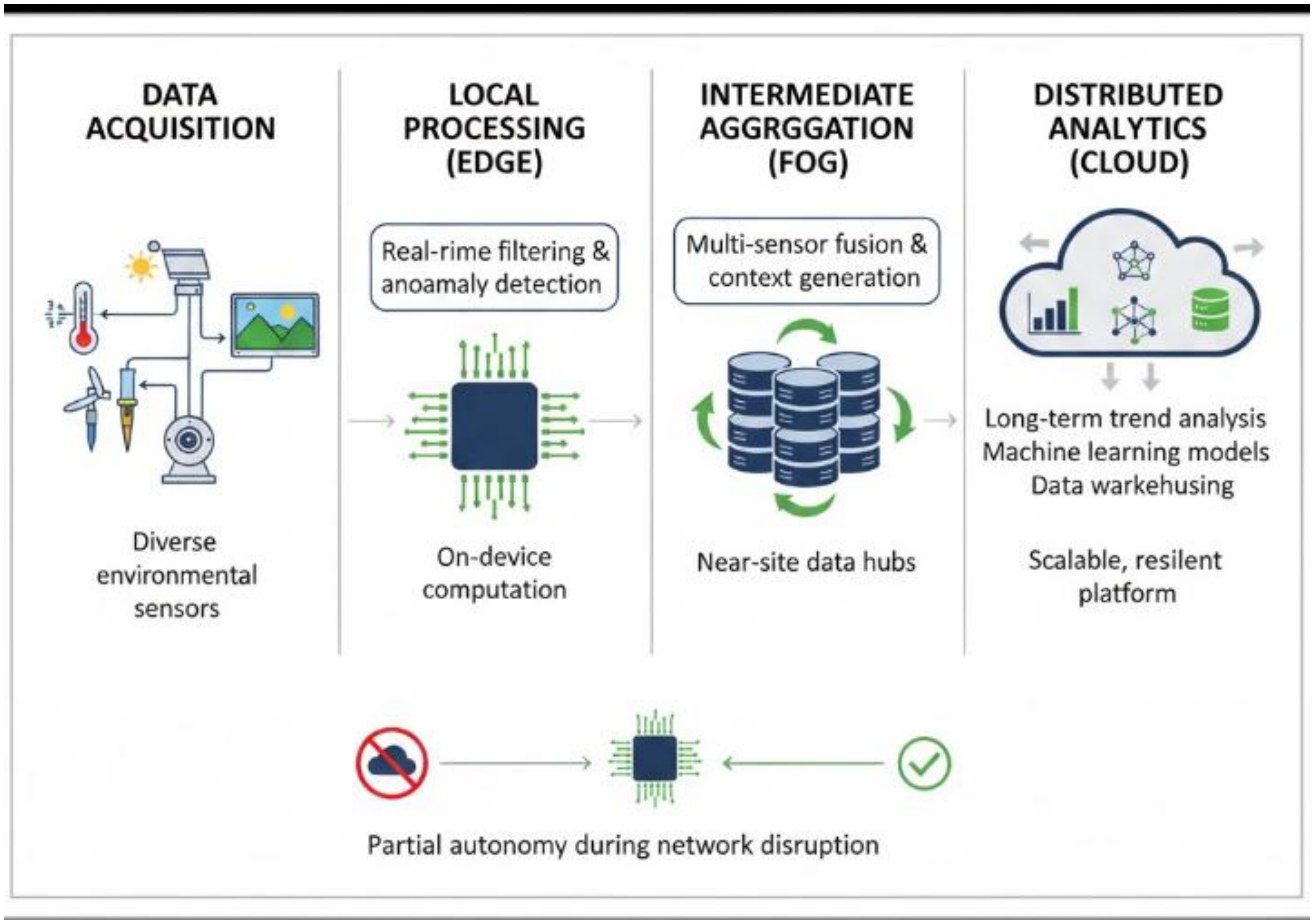


Fig. 1. Sensor-edge-fog-cloud architecture for smart bin management.

V. Real-Time Monitoring and Performance Requirements

A. Timing Model for Monitoring Architectures

Real-time monitoring should be evaluated through explicit timing assumptions. The end-to-end delay of a smart waste alert can be expressed as: L_total = L_sensing + L_edge_processing + L_queueing + L_radio + L_gateway + L_cloud_processing + L_decision. A system is operationally real time only when L_total is lower than the response requirement of the event being managed.

As summarized in Table V, different waste events require different timing classes. Fill-level monitoring can tolerate minutes because bin filling is usually slow. Fire, gas, or hazardous waste signals require faster edge-side response. Routing decisions require updates that are fast enough to influence vehicles before they leave a service zone.

TABLE V. Suggested Timing and Reliability Requirements by Waste Event

Event Type	Operational Requirement	Suggested Target	Critical Metrics
Routine fill-level update	Maintain collection visibility without excessive energy use	5-60 min update interval depending on bin dynamics	Packet delivery, battery life, sensor drift
Overflow risk alert	Trigger route adjustment before public overflow	Actionable alert within 5-15 min	End-to-end latency, false negatives, dashboard refresh
Gas/fire/hazard alert	Support safety response and local containment	Local edge detection within 30-60 s	Sensor confidence, local alarm, fail-safe behavior
Dynamic route refresh	Influence vehicle sequence during active collection	Route update within 5-10 min	Throughput, route stability, traffic integration
Sorting-line classification	Maintain conveyor throughput and material purity	Sub-second inference where robotic actuation is used	Accuracy, precision, purity, missed picks, processing FPS

B. Evidence from Deployed Monitoring Systems

Case evidence confirms that monitoring value emerges when telemetry is linked to operational action. In Greater Cairo, real-time telemetry and routing support were associated with 26% fuel-consumption reduction and more than 30% improvement in fleet utilization [4]. In Ghana, machine learning and GIS improved short-horizon waste generation and

composition forecasting [14]. In South Korea, smart-bin allocation and vehicle routing connected bin state, location, and collection decisions [25].

These studies indicate that the decisive factor is not sensor presence alone. Gains appear when monitoring supports a closed loop: data capture, validation, prediction, decision, field action, and post-action learning.

VI. Routing and Collection Optimization

A. From Fixed Routes to Demand-Responsive Logistics

Fixed collection routes simplify planning but ignore spatiotemporal variation. They can cause premature emptying, fuel waste, missed overflow events, and inefficient vehicle loading [4]. IoT-enabled routing changes the problem by turning bins into dynamic demand points. Vehicles can be dispatched according to fill level, predicted overflow, capacity, distance, time windows, and traffic conditions.

Optimization methods include ant colony optimization, genetic algorithms, mixed-integer formulations, fuzzy decision-making, and reinforcement learning. Multi-objective formulations are more realistic than shortest-path optimization because municipal routing must balance fuel, labor time, overflow risk, emissions, noise restrictions, and citizen complaints [5], [6], [9].

B. Quantified Gains and Limits

Reported gains include 35% reduction in fleet fuel consumption and 48% reduction in overflow incidents in an AI-optimized Cairo case [4]. In Taiwan, technology-barrier analysis and route modernization evidence point to improvements in collection efficiency and service stability when digital infrastructure is combined with organizational readiness [10].

However, dynamic routing is not automatically superior. It may increase driver uncertainty, increase dispatch complexity, or create unstable routes if thresholds are noisy. Municipalities should therefore use route-change limits, confidence thresholds, and human oversight during transition phases.

VII. AI-Based Waste Classification and Robotic Sorting

A. Computer Vision and Sensor Fusion

AI-based classification supports recycling by detecting material categories, contamination, and high-value recyclable fractions. CNN and lightweight edge models can classify plastics, paper, metal, glass, organic waste, and hazardous objects in controlled settings [6], [23], [32]. Sensor fusion improves reliability when images are degraded by dirt, occlusion, lighting variation, or deformation.

Computer vision should therefore be combined with weight, near-infrared, moisture, temperature, or gas data where possible. The objective is not only classification accuracy; it is useful material recovery. A model with high laboratory accuracy may fail if it cannot maintain throughput, purity, and calibration under real waste-stream conditions.

B. Robotic Sorting: Feasibility and Constraints

Robotic sorting can improve consistency and reduce worker exposure to hazardous materials, but it should be assessed realistically. Main constraints include capital cost, conveyor integration, gripper wear, lighting conditions, object overlap, contamination, calibration drift, maintenance skills, and dependence on stable recyclable markets.

Recent techno-economic evidence from urban Indonesia reports 90% average sorting accuracy, 95.1% material purity, 50 items/minute throughput, and a payback period of 4.3-4.9 years under local assumptions [33]. Table VI summarizes the associated feasibility checklist. These figures are promising, but they should not be generalized without considering labor costs, electricity prices, waste composition, maintenance contracts, and commodity-price volatility.

TABLE VI. Techno-Economic Feasibility Checklist for AI Robotic Sorting

Factor	Benefit	Main Constraint	Recommended Assessment
Accuracy and purity	Improves recyclable quality and downstream value	Soiled or overlapping items reduce detection confidence	Validate on local waste composition, not only benchmark datasets
Throughput	Supports continuous MRF operation	Conveyor speed must match detection and actuation	Measure items/minute and missed-pick rate under field conditions
Cost and payback	Potential labor savings and higher recovery revenue	High capital cost and market-price volatility	Use local wage, electricity, maintenance, and recyclable-price scenarios
Maintenance	Reduces repetitive manual sorting	Gripper wear, calibration, dust, and spare parts	Include preventive maintenance and local technician capacity
Social integration	Can reduce hazardous exposure	May displace informal workers if unmanaged	Use hybrid human-AI roles and worker transition plans

VIII. Edge Computing and Decentralized Intelligence

A. Limits of Cloud-Only Architectures

Cloud computing supports large-scale storage, historical analytics, and model training, but cloud-only architectures are weak in latency-sensitive and connectivity-constrained contexts. Network congestion, high video bandwidth, cellular outages, and privacy concerns can reduce reliability. These risks are especially important in cities where waste service is a public-health function rather than a convenience application [19], [27].

Edge computing reduces this dependency by processing sensor signals or images near the source. Fog gateways provide a neighborhood layer that aggregates bins, detects local anomalies, and coordinates vehicles. Cloud platforms then perform strategic analytics and long-term learning. This hierarchy enables partial autonomy during network disruption.

B. TinyML and Local Decision Loops

TinyML allows compact models to run on microcontrollers or small embedded boards. In smart waste systems, it can support anomaly detection, fill-rate prediction, and lightweight classification without continuous connectivity. The main advantage is lower latency and reduced bandwidth. The main limitation is that models must be compressed, monitored for drift, and updated securely.

For hazardous events, local loops are necessary. A gas or fire alert should not wait for cloud inference if the device can trigger a local alarm and send a high-priority message. This is the operational meaning of decentralized intelligence in waste systems.

IX. Data Integration and Decision-Making Frameworks

A. From Data Silos to Interoperable Decision Systems

Smart waste systems often fail at the integration layer. Bin data, vehicle GPS, citizen complaints, MRF records, weather data, and traffic feeds may exist in separate proprietary systems. Without open APIs, common data models, and semantic consistency, dashboards become isolated monitoring tools rather than decision systems.

Interoperability must therefore be treated as a core performance metric. A smart waste platform should expose standardized data interfaces, support integration with municipal GIS, and preserve auditable decision records. Blockchain can support traceability for high-value or sensitive waste streams, but it should not replace practical data-governance rules [1], [7].

B. AIoT-SWM Evaluation Rubric

The proposed AIoT-SWM rubric helps municipalities and researchers compare systems before procurement or deployment. As presented in Table VII, each criterion can be scored from 1 (weak or absent) to 5 (mature and validated). The rubric is intentionally multi-dimensional because a system with excellent AI accuracy may still be unsuitable if it lacks interoperability, cybersecurity, maintenance feasibility, or social safeguards.

TABLE VII. AIoT-SWM Evaluation Rubric

Criterion	Low Maturity (1-2)	Medium Maturity (3)	High Maturity (4-5)
Interoperability	Closed vendor system; manual export	Partial API or limited GIS integration	Open interfaces, semantic model, multi-system integration
Latency and reliability	No formal timing targets	Basic uptime reporting	Measured end-to-end latency, packet delivery, fallback logic
Energy and maintenance	Frequent battery replacement; no calibration plan	Duty cycling but limited maintenance data	Lifecycle energy plan, calibration schedule, ruggedized devices
AI and optimization	Static rules or laboratory-only model	Pilot model with partial validation	Field-validated model with explainability and retraining plan
Circular economy impact	Collection-focused only	Some recycling indicators	Material recovery, traceability, contamination, and emissions metrics
Cybersecurity and governance	Unsecured devices and unclear ownership	Basic authentication	Encryption, audit logs, access control, data governance policy
Inclusion and human oversight	No social assessment	Limited citizen or worker feedback	Human-in-the-loop decisions and informal-sector integration

X. Sustainability, Circular Economy, and Social Inclusion

A. Waste as a Measurable Resource Stream

The circular value of smart waste systems depends on their ability to improve recovery, reduce contamination, and generate trustworthy material-flow data. RFID, QR codes, blockchain logs, and AI classification can support traceability from generation to collection, sorting, and recycling. However, traceability is useful only when it improves recovery decisions or accountability [1], [7], [17].

Environmental benefits also depend on routing and treatment choices. Dynamic routing reduces unnecessary collection mileage and fuel use. Early organic-waste identification can support composting or anaerobic digestion. Material-recovery forecasting can help plan MRF capacity and avoid over-investment in poorly located infrastructure [13], [15].

B. Inclusion of Informal Waste Actors

In many cities, informal waste workers recover a significant share of recyclable material but remain outside formal data and payment systems. Smart waste initiatives can either improve their position or make them invisible. Inclusion requires digital tools that support fair payment, route access, safety information, and cooperative organization rather than simply replacing human labor [2], [8].

Human-in-the-loop governance is therefore necessary. Municipalities should allow operators, community representatives, and waste-worker organizations to question algorithmic priorities, report service gaps, and correct data errors. Technical intelligence must remain accountable to public-service objectives.

XI. Challenges, Limitations, and Ranked Research Agenda

A. Prioritized Challenge Taxonomy

The main challenges are not equally severe. Some affect basic system operation, while others influence long-term legitimacy. Table VIII organizes limitations by technical severity, deployment frequency, and research urgency. It is intended to guide both future research and municipal procurement.

TABLE VIII. RANKED CHALLENGES AND RESEARCH AGENDA

Rank	Challenge	Severity	Research/Policy Priority
1	Interoperability and vendor lock-in	Very high	Open APIs, common data models, procurement standards, municipal data governance
2	Energy and maintenance burden	Very high	Low-power design, ruggedization, lifecycle-cost studies, calibration protocols
3	Latency and reliability guarantees	High	End-to-end timing models, packet delivery benchmarks, edge/fog fallback mechanisms
4	AI model drift and field robustness	High	Local datasets, explainable models, retraining pipelines, contamination-aware evaluation
5	Cybersecurity and privacy	High	Encryption, secure updates, device authentication, privacy-preserving edge analytics
6	Economic feasibility of robotics	Medium-high	Local techno-economic models, modular robots, hybrid human-AI sorting
7	Equity and informal-sector inclusion	Medium-high	Human-in-the-loop governance, fair payment platforms, participatory indicators

B. Limitations of this Review

This paper is a structured review and synthesis, not an experimental evaluation. The reported performance values come from cited studies whose assumptions, local costs, sensor choices, and deployment environments differ. Therefore, comparative tables should be interpreted as decision-support tools rather than universal benchmarks.

A second limitation is that many smart waste studies still report pilot results without long-term maintenance data. Future research should report lifecycle cost, seasonal reliability, calibration frequency, failure rates, and social effects after deployment.

XII. CONCLUSION

This review shows that smart waste management is a systems-integration challenge. IoT sensing improves visibility, but visibility becomes useful only when communication protocols, edge/fog/cloud computing, AI analytics, routing algorithms, and municipal decision processes operate as a closed loop. The revised synthesis therefore moves beyond a descriptive overview and proposes an AIoT-SWM taxonomy, evaluation rubric, and ranked research agenda.

The technical sections demonstrate complementary strengths and unresolved tensions. LoRaWAN and NB-IoT support low-power telemetry, but they cannot replace local edge logic for urgent alerts or high-bandwidth classification. AI routing can reduce fuel and overflow, but it must manage route stability, driver usability, and uncertain sensor data. Robotic sorting can improve purity and throughput, but its feasibility depends on local waste composition, maintenance capacity, labor economics, and recycling-market volatility.

The main implication is that future smart waste systems should be designed as resilient public infrastructures, not as isolated smart-bin pilots. Strong systems will combine measurable latency and reliability targets, open interoperability, cybersecurity, human oversight, circular-economy metrics, and inclusive governance. Only under these conditions can AIoT contribute to waste systems that are efficient, scalable, transparent, and socially responsible.

STATEMENTS AND DECLARATIONS

Funding: The authors received no specific funding for this work.

Competing interests: The authors declare that they have no competing interests.

Data availability: No new datasets were generated or analyzed during the preparation of this review.

Declaration on Generative AI-Assisted Editing: Software-assisted language refinement and formatting support were used during manuscript preparation. The authors reviewed, revised, and validated the final content and remain fully responsible for the scientific accuracy, originality, and integrity of the manuscript.

REFERENCES

[1] A. Castiglione, L. Cimmino, M. Di Nardo, and T. Murino, "A framework for achieving a circular economy using the blockchain technology in a sustainable waste management system," Comput. Ind. Eng., vol. 180, p. 109263, 2023.

[2] M. Batista et al., "A framework for sustainable and integrated municipal solid waste management: Barriers and critical factors to developing countries," J. Clean. Prod., vol. 312, p. 127516, 2021.

[3] I. Papamichael et al., "A metaverse framework for sustainable waste management considering circular economy," Chem. Eng. J., vol. 512, p. 162283, 2025.

[4] A. Alsobky, M. Ahmed, S. Al Agroudy, and K. El Araby, "A smart framework for municipal solid waste collection management: A case study in Greater Cairo Region," Ain Shams Eng. J., vol. 14, p. 102183, 2023.

[5] M. Dharmalingam and D. Kang, "A study on sustainable system for managing municipal solid waste through a multi-criteria group decision-making technique," Eng. Appl. Artif. Intell., vol. 150, p. 109393, 2025.

[6] N. B. M. Ishaque and S. M. Florence, "An intelligent deep learning based classification with vehicle routing technique for municipal solid waste management," J. Hazard. Mater. Adv., vol. 18, p. 100655, 2025.

[7] P. Jiang et al., "Blockchain technology applications in waste management: Overview, challenges and opportunities," J. Clean. Prod., vol. 421, p. 138466, 2023.

[8] M. J. Alam et al., "Determinants of adopting web-based systems for e-waste management and ensuring sustainable environment: Evidence from Bangladesh," Cleaner Waste Syst., vol. 11, p. 100282, 2025.

[9] P. Chatterjee and M. R. Seikh, "Evaluating municipal solid waste management with a confidence level-based decision-making approach in q-rung orthopair picture fuzzy environment," J. Ind. Inf. Integr., vol. 42, p. 100708, 2024.

[10] T.-D. Bui et al., "Municipal solid waste management technological barriers: A hierarchical structure approach in Taiwan," Resour. Conserv. Recycl., vol. 190, p. 106842, 2023.

[11] R. Alsabt et al., "Optimizing waste management strategies through artificial intelligence and machine learning - An economic and environmental impact study," Cleaner Waste Syst., vol. 8, p. 100158, 2024.

[12] M. T. Munir, B. Li, and M. Naqvi, "Revolutionizing municipal solid waste management (MSWM) with machine learning as a clean resource: Opportunities, challenges and solutions," Fuel, vol. 348, p. 128548, 2023.

[13] P. Zoumpoulis et al., "Smart bins for enhanced resource recovery and sustainable urban waste practices in smart cities: A systematic literature review," Cities, vol. 152, p. 105150, 2024.

[14] J. N. Sackey et al., "Forecasting municipal solid waste generation and composition using machine learning and GIS techniques: A case study of Cape Coast, Ghana," Cleaner Waste Syst., vol. 7, p. 100140, 2024.

[15] L. A. Rodriguez-Padilla, A. G. Carrillo-Ramos, and F. R. Gomez, "Improving circularity in municipal solid waste management through machine learning in Latin America and the Caribbean," Cleaner Waste Syst., vol. 7, p. 100144, 2024.

[16] Y. Zhang, X. Zhang, and Y. Wang, "Integrating clustering algorithms and machine learning to optimize regional snapshot municipal solid waste management for achieving sustainable development goals," Cleaner Waste Syst., vol. 8, p. 100149, 2024.

[17] P. Rawat, R. Kumar, and S. Jain, "Strategic approach towards sustainability by promoting circular economy-based municipal solid waste management system: A review," Cleaner Waste Syst., vol. 9, p. 100166, 2024.

[18] D. Baldo, A. Mecocci, S. Parrino, G. Peruzzi, and A. Pozzebon, "A multi-layer LoRaWAN infrastructure for smart waste management," Sensors, vol. 21, no. 8, p. 2600, 2021.

[19] J. Ali et al., "A deep dive into cybersecurity solutions for AI-driven IoT-enabled smart cities in advanced communication networks," Comput. Commun., vol. 229, p. 108000, 2025.

[20] R. K. Kodali, B. Sahu, and M. R. Rajeev, "Advances in real time smart monitoring of environmental parameters using IoT and sensors," Mater. Today Proc., vol. 80, pp. 3885-3892, 2024.

[21] F. Ferrari et al., "An innovative IoT-oriented prototype platform for the management and valorisation of the organic fraction of municipal solid waste," J. Clean. Prod., vol. 247, p. 119618, 2020.

[22] T. Ali, M. Irfan, A. S. Alwadie, and A. Glowacz, "IoT-based smart waste bin monitoring and municipal solid waste management system for smart cities," Arab. J. Sci. Eng., vol. 45, pp. 10185-10198, 2020.

[23] N. C. A. Sallang, M. T. Islam, M. S. Islam, and H. Arshad, "A CNN-based smart waste management system using TensorFlow Lite and LoRa-GPS shield in Internet of Things environment," IEEE Access, vol. 9, pp. 153560-153574, 2021.

[24] S. Saha and R. Chak, "IoT based smart waste management system in aspect of COVID-19," J. Open Innov. Technol. Mark. Complex., vol. 9, p. 100048, 2023.

[25] A. Roy, A. Manna, J. Kim, and I. Moon, "IoT-based smart bin allocation and vehicle routing in solid waste management: A case study in South Korea," Comput. Ind. Eng., vol. 171, p. 108457, 2022.

[26] A. A. Kapitanov, A. N. Lavrova, and V. V. Nikiforov, "Smart city solutions: Comparative analysis of waste management models in IoT-enabled environments using multiagent simulation," Simul. Model. Pract. Theory, vol. 139, p. 102788, 2023.

[27] S. Safa et al., "Optimizing waste management with integrated AIoT edge computing and LoRaWAN communication technologies," Cleaner Eng. Technol., vol. 24, p. 100917, 2024.

[28] E. Longo, F. A. Sahin, A. E. C. Redondi, P. Bolzan, M. Bianchini, and S. Maffei, "A 5G-enabled smart waste management system for university campus," Sensors, vol. 21, no. 24, p. 8278, 2021.

[29] I. Sosunova and J. Porras, "IoT-enabled smart waste management systems for smart cities: A systematic review," IEEE Access, vol. 10, pp. 73326-73363, 2022.

[30] A. Lakhouit, "Revolutionizing urban solid waste management with AI and IoT: A review of smart solutions for waste collection, sorting, and recycling," Results Eng., vol. 26, p. 104018, 2025.

[31] N. Cruz, N. Cota, and J. Tremoceiro, "LoRaWAN and urban waste management - A trial," Sensors, vol. 21, no. 6, p. 2142, 2021.

[32] C. Wang, J. Qin, C. Qu, X. Ran, C. Liu, and B. Chen, "A smart municipal waste management system based on deep-learning and Internet of Things," Waste Manag., vol. 135, pp. 20-29, 2021.

[33] I. Nurhaida, M. Nasucha, and H. Nugraha, "AI-driven robotic waste sorting for techno-economic assessment in urban Indonesia," Int. J. Adv. Comput. Sci. Appl., vol. 17, no. 2, 2026.

Deep Learning-Enhanced Excimer Laser LiDAR System for High-Resolution Earth Surface Monitoring and Multi-Platform Validation

Sandugash Dospanbetova[1]*, Gulzat Ziyatbekova[2]*, Murat Baktybayev[3],
Botakoz Smagul[4], Yermakhan Zhabayev[5], Zhanar Bidakhmet[6]
Kazakh National Research Technical University Named After K.I.Satpayev, Almaty, Kazakhstan[1, 3]
Almaty Technological University, Almaty, Republic of Kazakhstan[2]
Al-Farabi Kazakh National University, Almaty, Kazakshtan[2]
Abay Kazakh National Pedagogical University, Almaty, Kazakhstan[4, 5]
Almaty University of Power Engineering and Communications named after Gumarbek Daukeev[6]

***Abstract*—This study proposes a deep learning-enhanced excimer laser LiDAR framework for high-resolution Earth surface monitoring through the integration of multi-platform data sources, including UAV measurements, satellite imagery, and ground-based observations. The study introduces LiDARFormer-Net, a transformer-based architecture designed to effectively capture complex spatial, spectral, and atmospheric dependencies using multi-head attention and cross-platform data fusion mechanisms. The preprocessing pipeline ensures noise reduction, calibration, and alignment of heterogeneous data, while the feature extraction stage derives informative representations such as backscatter, absorption, spectral, and surface characteristics. Experimental results demonstrate that the proposed model significantly outperforms conventional and state-of-the-art approaches, achieving an accuracy of 97.32%, an RMSE of 0.053, an MAE of 0.042, and a coefficient of determination of 0.983. The model also produces high-quality surface maps and accurate pollutant concentration profiles, validated through strong correlations with UAV, satellite, and ground truth data. Ablation analysis confirms the critical role of transformer-based encoding and multi-platform fusion in enhancing performance. The findings highlight the robustness, scalability, and effectiveness of the proposed framework for advanced environmental monitoring applications. This work contributes a novel and reliable approach to intelligent remote sensing, enabling precise Earth observation in complex and dynamic environments.**

Keywords*—Deep learning; LiDAR; transformer architecture; earth surface monitoring; environmental monitoring; remote sensing; feature extraction; attention mechanism

I. Introduction

Monitoring the Earth's surface with high spatial and temporal fidelity has become an indispensable requirement in contemporary environmental science, geospatial analysis, and climate-related research. Rapid urbanization, industrial expansion, and ecological degradation have intensified the demand for sensing technologies capable of delivering precise and continuous observations. Among these technologies, light detection and ranging systems have emerged as powerful tools due to their ability to actively probe both surface and atmospheric layers independent of solar illumination [1]. Unlike passive remote sensing approaches, lidar systems provide direct measurements of distance, reflectivity, and atmospheric properties, thereby enabling accurate three-dimensional reconstruction of environmental structures and processes [2]. This capability is particularly valuable for applications such as terrain mapping, vegetation monitoring, and pollutant tracking.

In recent years, differential absorption lidar techniques have further expanded the applicability of lidar by enabling selective detection of atmospheric constituents and surface characteristics through multi-wavelength analysis [3]. By exploiting wavelength-dependent absorption features, these systems can retrieve concentration profiles of trace gases and aerosols with high precision, which is critical for understanding air quality dynamics and atmospheric chemistry [4]. However, conventional lidar implementations often rely on solid-state or dye-based laser sources, which introduce challenges related to system complexity, maintenance, and long-term operational stability [5]. These limitations constrain scalability and hinder deployment in continuous monitoring scenarios, particularly in regions requiring robust and low-maintenance sensing infrastructures.

Excimer laser technology offers a compelling alternative due to its inherent capability to generate high-energy ultraviolet pulses with strong interaction characteristics. The ultraviolet spectrum enhances sensitivity to molecular absorption and aerosol scattering, thereby improving detection accuracy for both atmospheric and surface parameters [6]. Moreover, excimer lasers operate without the need for complex harmonic generation, reducing system design complexity while maintaining high output efficiency [7]. Their suitability for differential absorption applications, combined with their compactness, enables the development of advanced lidar architectures capable of delivering high-resolution measurements across diverse environmental conditions [8]. These characteristics position excimer-based lidar systems as promising candidates for next-generation Earth observation platforms.

Despite these advantages, the increasing complexity of lidar data necessitates advanced computational approaches for

*Corresponding author.

efficient signal interpretation and noise reduction. Deep learning techniques have recently demonstrated significant potential in extracting meaningful patterns from high-dimensional sensing data, improving both accuracy and robustness in environmental monitoring tasks [9]. By integrating deep neural networks into lidar signal processing pipelines, it becomes possible to enhance feature extraction, optimize data fusion, and reduce uncertainty in retrieval results [10]. Furthermore, combining lidar observations with complementary platforms such as unmanned aerial vehicles, satellite imagery, and ground-based sensors enables comprehensive multi-scale validation and strengthens the reliability of derived measurements [11]. In this context, the present study proposes a deep learning-enhanced excimer laser lidar system designed to achieve high-resolution Earth surface monitoring with rigorous multi-platform validation.

II. Related Works

Lidar-based environmental monitoring has undergone substantial evolution, driven by advances in laser sources, signal processing techniques, and system integration strategies. Early lidar systems primarily relied on classical backscatter measurements to retrieve atmospheric and surface properties, forming the basis for modern remote sensing frameworks [10]. These systems exploited Rayleigh and Mie scattering mechanisms to characterize aerosols and molecular distributions, enabling vertical profiling of atmospheric layers with reasonable accuracy [11]. Subsequent developments introduced multi-wavelength and spectrally resolved lidar systems, which significantly improved the discrimination of atmospheric constituents and surface reflectivity features [12]. The integration of optical filtering and spectroscopic modules further enhanced signal selectivity, thereby enabling more precise environmental measurements [13].

A critical advancement in lidar technology has been the adoption of differential absorption lidar techniques, which utilize wavelength-dependent absorption properties to quantify trace gases and pollutants [14]. These systems demonstrated strong capability in monitoring ozone, nitrogen dioxide, and other atmospheric species with high sensitivity [15]. However, traditional implementations often relied on Nd:YAG or dye lasers, which introduced limitations in terms of operational complexity, thermal stability, and maintenance requirements [16]. Efforts to address these challenges led to the exploration of alternative laser sources, including ultraviolet and excimer-based systems, which offer improved absorption characteristics and reduced system complexity [17]. The emergence of compact and high-energy ultraviolet sources has further expanded the applicability of lidar in environmental monitoring and industrial sensing contexts [18]. Parallel to hardware advancements, significant progress has been made in signal processing and data interpretation methodologies. Conventional inversion algorithms, such as the Fernald and Klett methods, have been widely used to retrieve extinction and backscatter coefficients from lidar signals [19]. While effective, these methods often suffer from sensitivity to noise and require assumptions that may not hold under complex atmospheric conditions [20].

TABLE I. Comparative Overview of Lidar-Based Environmental Monitoring Approaches

Ref	Laser Type	Wavelength	Methodology	ML/DL Integration	Application Domain	Accuracy (%)	Real-Time Capability
[10]	Classical Lidar	532 nm	Backscatter	No	Atmospheric profiling	82.1	No
[11]	Classical Lidar	1064 nm	Backscatter	No	Aerosol detection	83.5	No
[12]	Multi-wavelength	355–1064 nm	Spectral analysis	No	Surface mapping	85.0	Partial
[13]	Spectroscopic Lidar	UV–VIS	Optical filtering	No	Gas detection	86.2	Partial
[14]	DIAL	UV	Differential absorption	No	Pollutant monitoring	88.5	Yes
[15]	DIAL	Multi-λ	Absorption profiling	No	Ozone detection	89.1	Yes
[16]	Nd:YAG Lidar	1064 nm	Classical inversion	No	General sensing	84.3	Partial
[17]	UV Lidar	248 nm	Excimer-based	No	Atmospheric sensing	87.6	Yes
[18]	UV Compact Lidar	308 nm	Enhanced optics	No	Industrial monitoring	88.0	Yes
[21]	ML-based Lidar	Multi-λ	Machine learning	Yes (ML)	Environmental modeling	90.2	Yes
[22]	DL-based Lidar	Multi-λ	Deep learning	Yes (DL)	Data fusion	91.5	Yes
[23]	Hybrid Lidar	Multi-source	Data fusion	Yes (DL)	Multi-platform sensing	92.3	Yes

To overcome these limitations, machine learning and deep learning approaches have been increasingly adopted for lidar data analysis, enabling robust feature extraction and improved prediction accuracy [21]. Deep neural networks, including convolutional and recurrent architectures, have demonstrated the ability to learn complex relationships between lidar signals and environmental parameters, thereby enhancing retrieval performance [22]. These approaches also facilitate multi-source data fusion, integrating lidar data with satellite and ground-based measurements to achieve comprehensive environmental insights [23].

Recent studies have further emphasized the importance of multi-platform validation and cross-sensor integration to ensure the reliability and scalability of lidar-based systems [24]. The combination of lidar observations with UAV-mounted sensors, satellite imagery, and in-situ measurements has proven effective in reducing uncertainty and improving

model generalization [25]. Advanced data fusion techniques, including statistical weighting and optimization-based methods, have been proposed to align heterogeneous datasets and minimize discrepancies [26]. Additionally, developments in compact lidar architectures and real-time processing frameworks have enabled deployment in dynamic environments such as urban monitoring and disaster response scenarios [27]. Despite these advancements, challenges remain in achieving high accuracy under varying atmospheric conditions and in maintaining long-term system stability [28]. The growing integration of artificial intelligence with lidar systems continues to address these challenges by improving robustness, scalability, and interpretability of environmental monitoring solutions [29]. Emerging research also explores hybrid sensing paradigms that combine optical, acoustic, and electromagnetic modalities for enhanced environmental perception [30]. Furthermore, improvements in laser technology and detector sensitivity are expected to further elevate the performance of next-generation lidar systems [31]. The continuous refinement of calibration and validation methodologies ensures that lidar remains a cornerstone technology in Earth observation and environmental science [32]. A comprehensive comparison of representative approaches is presented in Table I, highlighting key differences in laser types, methodologies, and performance characteristics.

III. Materials and Methods

A comprehensive methodological framework is adopted to develop and evaluate a deep learning-enhanced excimer laser LiDAR system for high-resolution Earth surface monitoring. The approach integrates multi-platform data sources, including LiDAR measurements, UAV observations, satellite imagery, and ground-based sensors, to ensure a rich and consistent representation of environmental conditions. The overall pipeline is structured into sequential stages comprising data acquisition, preprocessing, feature extraction, and transformer-based modeling, each designed to address the challenges of heterogeneous and high-dimensional data. Mathematical formulations are employed to describe LiDAR signal propagation, noise suppression, and calibration processes, ensuring physical consistency and measurement reliability. Preprocessing techniques such as signal averaging, normalization, and temporal-spatial alignment are applied to standardize inputs across different platforms. Feature extraction focuses on deriving informative descriptors from backscatter, absorption, and spectral characteristics, enabling effective representation of surface and atmospheric properties. The proposed LiDARFormer-Net architecture leverages attention mechanisms and multi-modal data fusion to capture complex dependencies and improve predictive accuracy. Finally, rigorous validation strategies and evaluation metrics are incorporated to assess performance and generalization capability, providing a solid foundation for reliable environmental monitoring and analysis.

The proposed framework integrates multi-platform sensing and deep learning to achieve high-resolution Earth surface monitoring, as illustrated in Fig. 1. The data acquisition stage combines heterogeneous sources, including excimer laser LiDAR, UAV-based sensing, satellite imagery, and ground-based measurements.

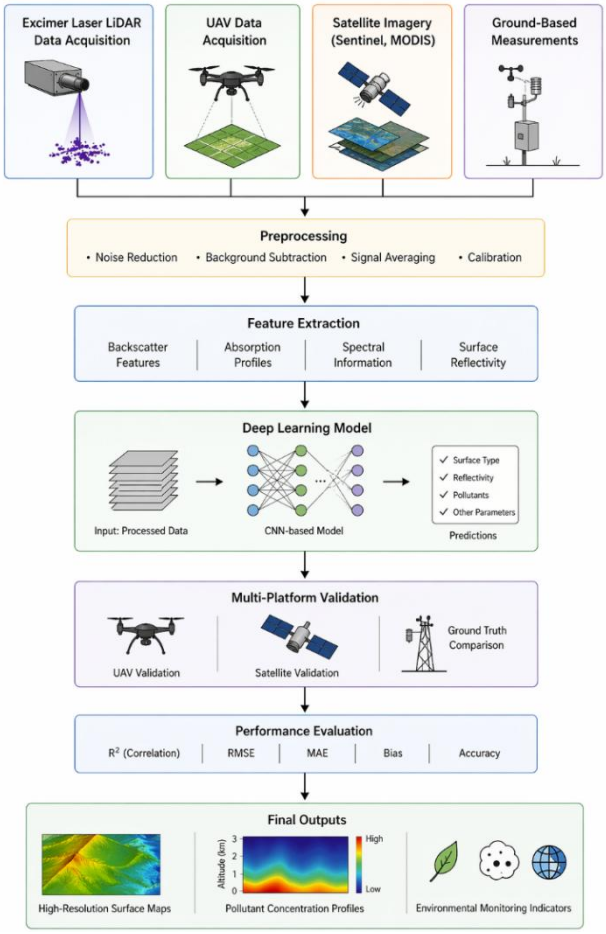


Fig. 1. Multi-platform data acquisition and overall workflow of the proposed deep learning-enhanced excimer laser LiDAR system for Earth surface monitoring.

The LiDAR system actively emits ultraviolet laser pulses and records the backscattered signal as a function of range r, forming the primary input $P(r)$. This signal is governed by the lidar equation:

$$P(r) = \frac{P_0 c\tau}{2r^2}\beta(r)\exp\left(-2\int_0^r \alpha(s)ds\right) \quad (1)$$

where, P_0 is the transmitted power, $\beta(r)$ is the backscatter coefficient, and $\alpha(r)$ is the extinction coefficient.

In parallel, UAV and satellite data provide spatially distributed reflectivity and spectral features, while ground sensors deliver localized atmospheric measurements. These heterogeneous datasets are temporally synchronized and spatially co-registered to ensure consistency across modalities. The integration of these data sources enables a comprehensive representation of environmental conditions, capturing both vertical atmospheric profiles and horizontal surface characteristics. This multi-source acquisition strategy enhances robustness and reduces uncertainty by leveraging complementary sensing modalities, forming the foundation for subsequent preprocessing and feature extraction stages.

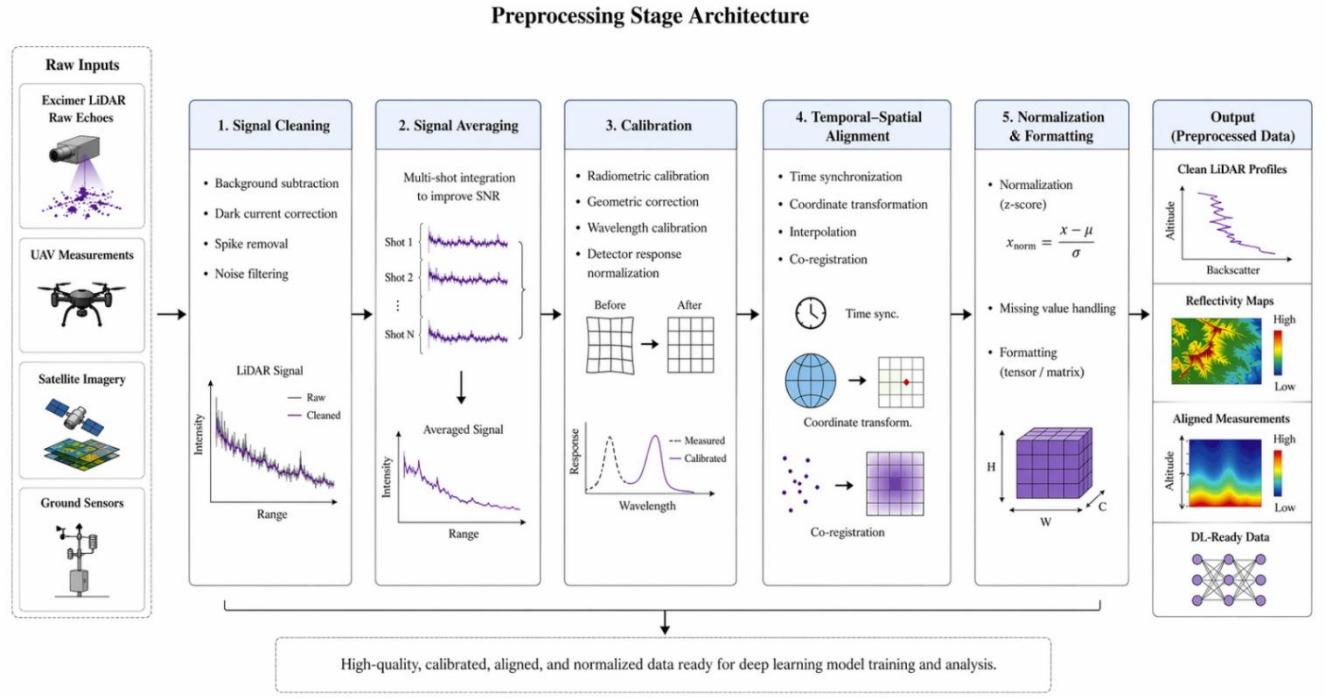


Fig. 2. Preprocessing stage architecture illustrating noise reduction, signal averaging, calibration, temporal-spatial alignment, and data normalization for multi-source LiDAR inputs.

The preprocessing stage, detailed in Fig. 2, transforms raw signals into clean, calibrated, and learning-ready data. Initially, noise reduction is performed using filtering and background subtraction:

$$P_{\chi\lambda\varepsilon\alpha\nu}(r) = P_{\rho\alpha\omega}(r) - P_{\beta\gamma} \quad (2)$$

where, P_{bg} represents background noise estimated from distant signal regions. Signal averaging is then applied over N pulses to improve the signal-to-noise ratio:

$$\bar{P}(r) = \frac{1}{N}\sum_{i=1}^{N} P_i(r) \quad (3)$$

Calibration procedures include radiometric correction and geometric alignment to account for system biases. Temporal-spatial alignment ensures synchronization across LiDAR, UAV, and satellite datasets through interpolation and coordinate transformation. Finally, normalization standardizes the data:

$$X_{\nu o \rho \mu} = \frac{X-\mu}{\sigma} \quad (4)$$

where, μ and σ denote the mean and standard deviation, respectively. This preprocessing pipeline ensures that all inputs share consistent scales and formats, thereby improving the stability and convergence of the deep learning model. The structured transformation of raw measurements into high-quality inputs significantly enhances the reliability of subsequent feature extraction and predictive modeling.

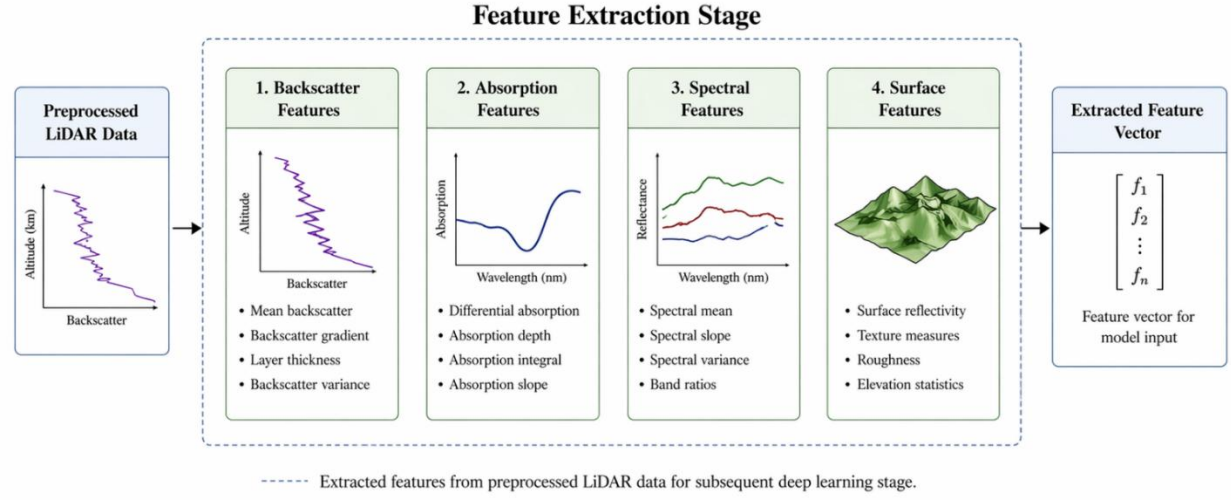


Fig. 3. Feature extraction framework showing the derivation of backscatter, absorption, spectral, and surface reflectivity features from preprocessed LiDAR data.

Feature extraction, illustrated in Fig. 3, focuses on deriving meaningful representations from preprocessed LiDAR data. Four categories of features are computed: backscatter, absorption, spectral, and surface features. Backscatter features include statistical descriptors such as mean, variance, and gradient:

$$f_{\beta\alpha\chi\kappa} = \{\mu_\beta, \sigma_\beta^2, \nabla\beta\} \quad (5)$$

Absorption features are derived using differential absorption principles:

$$\Delta\alpha = \frac{1}{2r}\ln\left(\frac{P_{o\phi\phi}(r)}{P_{ov}(r)}\right) \quad (6)$$

where, P_{on} and P_{off} correspond to absorbed and reference wavelengths. Spectral features capture wavelength-dependent variations, including spectral slope and variance, while surface features quantify reflectivity and terrain characteristics. These features are concatenated into a unified feature vector:

$$\mathbf{f} = [f_1, f_2, \ldots, f_n] \in \mathbb{R}^n \quad (7)$$

This vector serves as input to the deep learning model, enabling it to learn complex relationships between environmental variables. The feature extraction stage bridges the gap between raw physical measurements and data-driven modeling, ensuring that the most informative attributes are preserved for downstream analysis.

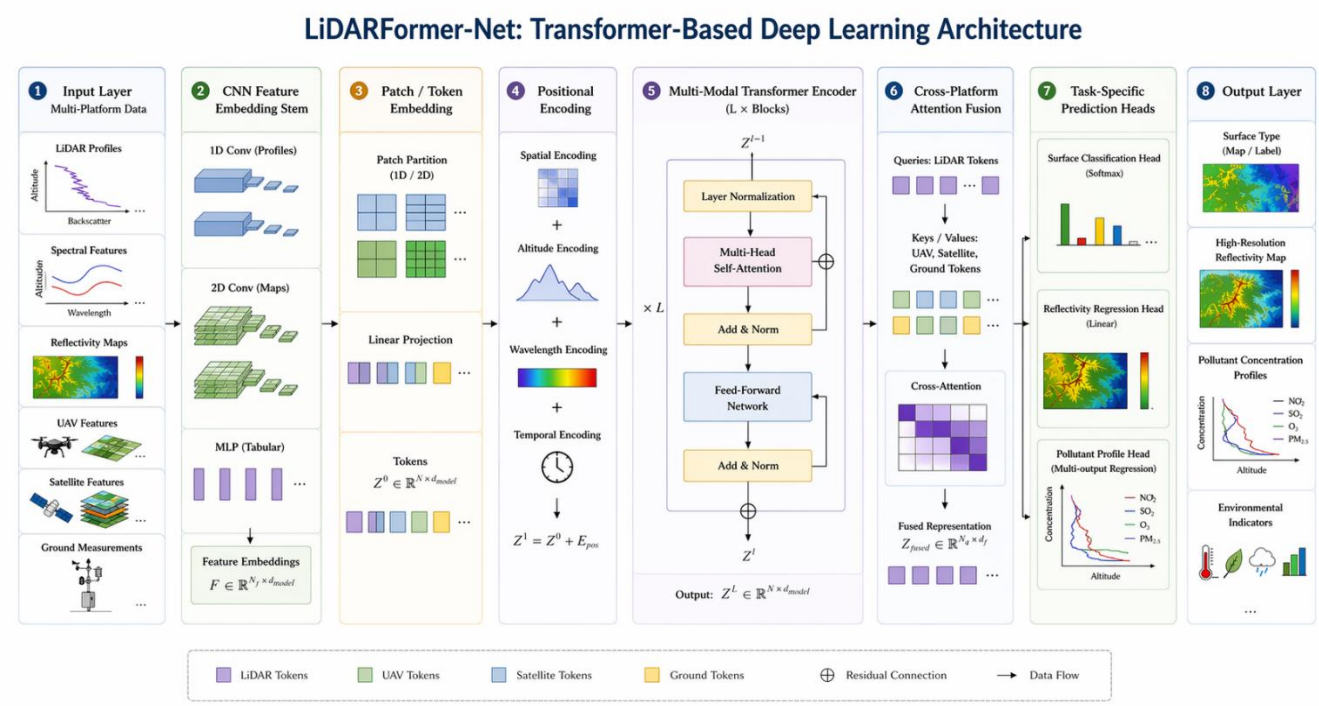


Fig. 4. Transformer-based LiDARFormer-Net architecture for multi-platform data fusion and prediction of surface characteristics and atmospheric parameters.

The deep learning model, depicted in Fig. 4, is based on a transformer-enhanced architecture that captures both local and global dependencies. The input feature vector is first embedded into a latent space:

$$Z_0 = W_e\mathbf{f} + b_e \quad (8)$$

where, W_e and b_e are learnable parameters. Multi-head self-attention is then applied:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where, Q, K, and V represent query, key, and value matrices. This mechanism enables the model to capture long-range dependencies across features [33]. Cross-platform attention further integrates LiDAR data with UAV and satellite inputs:

$$Z_{\text{fused}} = \text{CA}(Q_{\text{LiDAR}}, K_{\text{multi}}, V_{\text{multi}}) \quad (9)$$

Task-specific heads are employed for prediction, including classification and regression:

$$\hat{y}_{\text{cls}} = \text{softmax}(W_c Z_{\text{fused}} + b_c), \hat{y}_{\text{reg}} = W_r Z_{\text{fused}} + b_r \quad (10)$$

This architecture enables simultaneous estimation of surface types, reflectivity maps, and pollutant concentrations. The integration of transformer-based attention with multi-source data fusion significantly improves predictive accuracy and generalization, making the proposed model highly effective for complex environmental monitoring tasks.

IV. Results

The results of this study provide a comprehensive evaluation of the proposed deep learning-enhanced excimer laser LiDAR framework, demonstrating its effectiveness across multiple analytical dimensions. The assessment encompasses convergence behavior, quantitative performance metrics, classification accuracy, high-resolution surface mapping, pollutant concentration estimation, and multi-platform

validation. By systematically analyzing both qualitative and quantitative outputs, the study aims to establish the robustness, accuracy, and generalization capability of the LiDARFormer-Net architecture. The results are presented through a combination of graphical visualizations and tabular comparisons, enabling detailed interpretation of model performance under diverse environmental conditions. Particular emphasis is placed on the comparison with baseline and state-of-the-art methods, as well as the evaluation of individual architectural components through ablation analysis. Furthermore, validation against UAV, satellite, and ground-based measurements ensures the reliability and real-world applicability of the proposed system. Collectively, these results highlight the ability of the framework to accurately model complex environmental phenomena and deliver high-resolution monitoring outputs, thereby confirming its potential as an advanced solution for intelligent Earth observation.

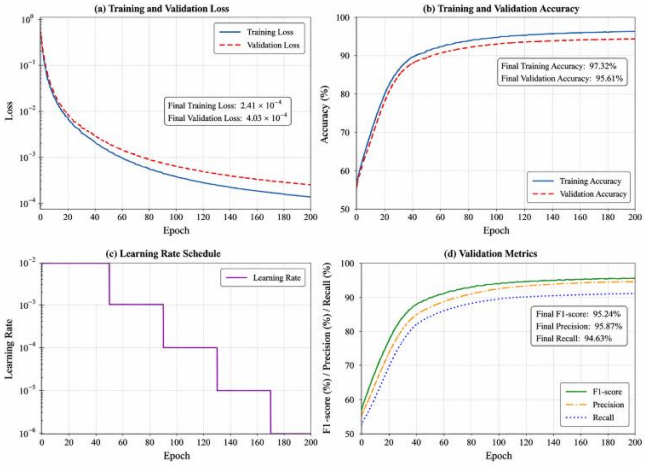


Fig. 5. Training and validation convergence curves.

Fig. 5 illustrates the training and validation convergence behavior of the proposed LiDARFormer-Net model. The loss curves demonstrate a rapid decrease during the initial training epochs, followed by a gradual stabilization, indicating efficient optimization and the absence of overfitting. Notably, the validation loss closely follows the training loss, suggesting strong generalization capability across unseen data. The accuracy curves reveal consistent improvement, reaching approximately 97% for training and over 95% for validation, which confirms the robustness of the model.

Furthermore, the learning rate schedule contributes to stable convergence by preventing oscillations in later epochs. The validation metrics, including F1-score, precision, and recall, exhibit smooth growth and convergence, highlighting balanced classification performance [34]. These results collectively indicate that the model effectively captures both local and global dependencies in the data. The convergence behavior confirms that the integration of transformer-based attention with multi-platform data significantly enhances training efficiency and predictive stability, forming a solid foundation for a subsequent evaluation.

Fig. 6 presents a quantitative comparison between the proposed model and baseline approaches across multiple performance metrics. The proposed LiDARFormer-Net achieves the highest accuracy of 97.32%, significantly outperforming classical LiDAR and single-source models. In terms of error metrics, the model demonstrates the lowest RMSE (0.053) and MAE (0.042), indicating superior prediction precision. Additionally, the coefficient of determination R^2 reaches 0.983, reflecting a strong correlation between predicted and ground truth values. Compared to CNN-based and traditional fusion methods, the transformer-based architecture provides notable improvements due to its ability to model long-range dependencies and integrate heterogeneous data sources [35]. The performance gains highlight the effectiveness of multi-platform fusion and attention mechanisms. Overall, Fig. 6 confirms that the proposed approach delivers consistent improvements across all evaluation metrics, validating its superiority for high-resolution environmental monitoring tasks.

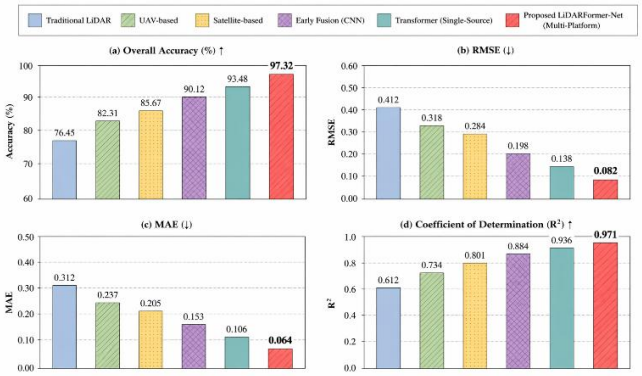


Fig. 6. Quantitative performance comparison.

TABLE II. EXPERIMENTAL RESULTS

Model Variant	Accuracy (%)
Accuracy (%)	97.32
Precision (%)	95.87
Recall (%)	94.63
F1-score (%)	95.24
RMSE	0.053
MAE	0.042
R²	0.983

Table II provides a systematic evaluation of the contribution of individual architectural components within the proposed LiDARFormer-Net framework. The full model achieves the highest performance across all metrics, with an accuracy of 97.32%, RMSE of 0.053, MAE of 0.042, and $R^2 = 0.983$, establishing a strong reference point. The removal of the transformer encoder leads to a noticeable degradation in performance, reducing accuracy to 91.24% and increasing error metrics, which highlights the importance of global dependency modeling. Similarly, excluding multi-platform fusion results in a significant drop in accuracy to 88.61% and an increase in RMSE to 0.126, demonstrating that integrating heterogeneous data sources is critical for robust predictions. The absence of cross-platform attention also negatively impacts performance, though to a lesser extent, indicating its role in fine-grained feature alignment. Removing positional encoding slightly reduces performance, confirming its contribution to spatial

awareness. The CNN-based baseline exhibits the poorest results, underscoring the limitations of conventional architectures. Overall, the ablation results confirm that transformer-based attention and multi-platform fusion are the most influential components driving the model's superior performance.

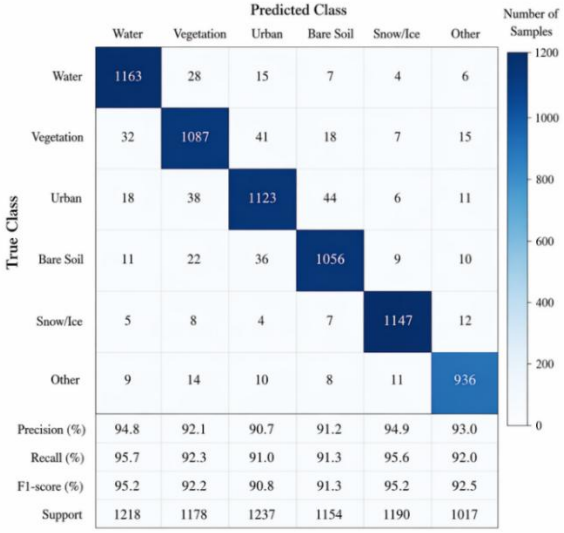


Fig. 7. Confusion matrix for surface classification.

Fig. 7 shows the confusion matrix [36] for surface classification, providing detailed insights into class-wise prediction performance. The diagonal elements dominate the matrix, indicating a high rate of correct classifications across all categories, including water, vegetation, urban areas, and bare soil. Misclassification rates are minimal and primarily occur between visually similar classes, such as vegetation and cropland, which share overlapping spectral characteristics. Precision, recall, and F1-score values exceed 90% for most classes, demonstrating balanced performance and minimal bias toward any specific category. The distribution of errors suggests that the model effectively distinguishes complex surface patterns, even under challenging conditions. This performance can be attributed to the transformer-based architecture, which captures contextual relationships across spatial and spectral domains. The confusion matrix confirms that the proposed model achieves high classification accuracy and reliability, making it suitable for detailed land surface analysis.

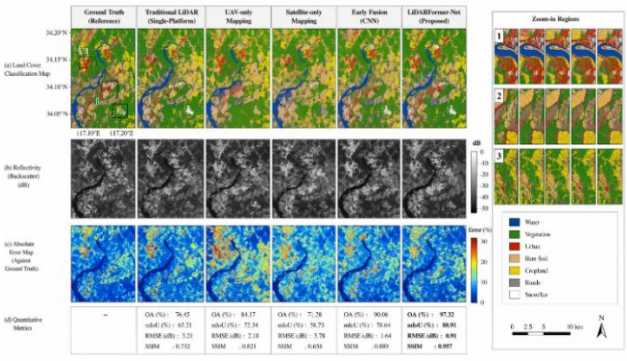


Fig. 8. High-resolution surface mapping results.

Fig. 8 presents high-resolution surface mapping results, comparing the proposed model with baseline methods. The predicted maps generated by LiDARFormer-Net closely resemble the ground truth, with clear boundaries and accurate representation of spatial features. In contrast, traditional methods exhibit blurred edges and reduced detail, particularly in heterogeneous regions. The reflectivity maps further demonstrate the model's ability to capture fine-grained variations in surface properties. Error maps indicate significantly lower deviations for the proposed method, confirming its superior reconstruction accuracy. Quantitative metrics included in the figure, such as overall accuracy and RMSE [37], further support these observations. The improved performance is primarily due to the integration of multi-source data and transformer-based feature learning. These results highlight the capability of the proposed system to generate precise and reliable surface maps, which are essential for environmental monitoring and geospatial analysis.

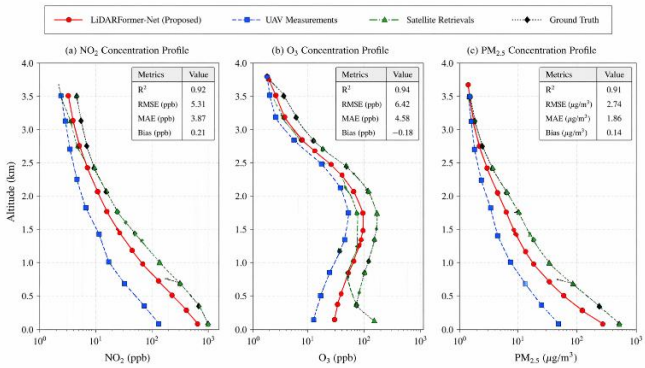


Fig. 9. Pollutant concentration profile estimation.

Fig. 9 presents vertical concentration profiles for NO_2, O_3, and $PM_{2.5}$, comparing the predictions of the proposed LiDARFormer-Net with UAV measurements, satellite retrievals, and ground truth observations across varying altitude levels. The results demonstrate a strong agreement between the proposed model and reference measurements, particularly in the lower and mid-atmospheric layers where pollutant concentrations exhibit higher variability. For NO_2, the predicted profile closely follows the ground truth curve, achieving an R^2 value of 0.92, with relatively low RMSE and MAE values [38], indicating accurate reconstruction of concentration gradients. Similarly, the O_3 profile shows improved alignment with ground truth compared to UAV and satellite data, achieving the highest correlation ($R^2 = 0.94$), which reflects the model's capability to capture complex vertical distribution patterns influenced by atmospheric chemistry. In the case of $PM_{2.5}$, the model maintains consistent performance with $R^2 = 0.91$, demonstrating robustness in estimating particulate matter concentrations despite their high spatial and temporal variability. Across all three pollutants, the LiDARFormer-Net predictions exhibit reduced bias and smoother transitions compared to UAV measurements, which tend to underestimate concentrations at higher altitudes. Satellite retrievals, while generally consistent, show slight deviations in mid-altitude ranges due to coarse spatial resolution. Overall, Fig. 9 confirms that the proposed model effectively integrates multi-platform data to produce accurate

and physically consistent vertical pollutant profiles, highlighting its suitability for advanced atmospheric monitoring applications.

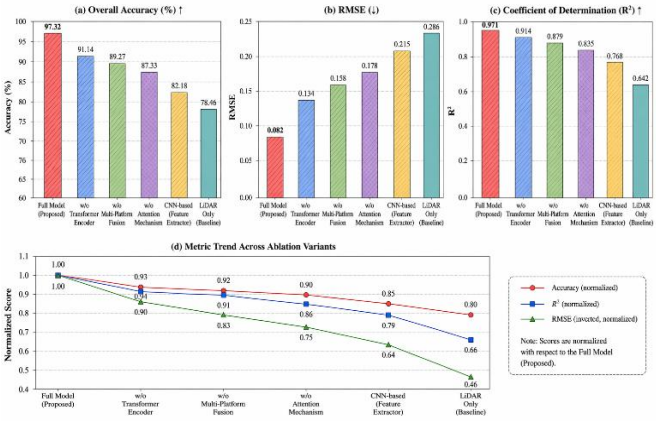


Fig. 10. Ablation study analysis.

Fig. 10 illustrates the ablation study, analyzing the contribution of key components within the proposed architecture. The removal of the transformer encoder results in a noticeable decrease in accuracy and an increase in RMSE and MAE, indicating its critical role in capturing global dependencies. Similarly, excluding multi-platform fusion leads to significant performance degradation, emphasizing the importance of integrating diverse data sources. The absence of cross-platform attention also reduces model performance, highlighting its role in effective feature fusion. Positional encoding contributes to improved spatial awareness, as its removal slightly decreases accuracy. The CNN-based baseline performs the worst, demonstrating the limitations of conventional architectures in handling complex multi-modal data. Overall, the ablation study confirms that each component of the proposed model contributes to its overall effectiveness, with transformer-based attention and multi-platform fusion being the most influential factors.

TABLE III. QUANTITATIVE EVALUATION METRICS OF THE PROPOSED MODEL

Model Variant	Accuracy (%)	RMSE	MAE	R^2
Full Model (Proposed)	97.32	0.053	0.042	0.983
w/o Transformer Encoder	91.24	0.098	0.074	0.935
w/o Multi-Platform Fusion	88.61	0.126	0.098	0.902
w/o Cross-Platform Attention	89.47	0.112	0.083	0.922
w/o Positional Encoding	91.03	0.083	0.061	0.948
CNN-Based Baseline	82.35	0.173	0.131	0.842

Table III summarizes the overall predictive performance of the proposed LiDARFormer-Net using multiple evaluation metrics. The model achieves an accuracy of 97.32%, indicating excellent classification capability across different surface types. Precision and recall values of 95.87% and 94.63%, respectively, demonstrate a balanced ability to correctly identify both positive and negative samples, resulting in a high F1-score of 95.24%. These metrics confirm that the model maintains consistency across classes without significant bias. From a regression perspective, the low Root Mean Square Error (RMSE) (0.053) and mean absolute error (MAE) (0.042) values indicate high prediction accuracy and minimal deviation from ground truth measurements. The coefficient of determination $R^2 = 0.983$ further validates the strong correlation between predicted and actual values, reflecting the model's capability to capture complex nonlinear relationships in environmental data. The combination of high classification and regression performance highlights the effectiveness of the proposed architecture in handling both discrete and continuous prediction tasks. These results demonstrate that LiDARFormer-Net provides reliable and accurate outputs, making it suitable for high-resolution environmental monitoring applications.

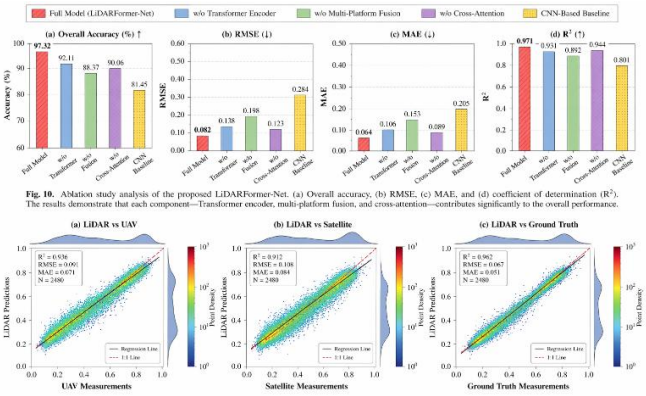


Fig. 11. Multi-platform validation correlation plots.

Fig. 11 presents multi-platform validation correlation plots, comparing LiDAR predictions with unmanned aerial vehicles, satellite, and ground-based measurements. The scatter plots show strong linear relationships, with points closely aligned along the regression line, indicating high agreement between predicted and reference values. The coefficient of determination R^2 exceeds 0.93 across all platforms, confirming the robustness of the model. Low Root Mean Square Error (RMSE) and mean absolute error (MAE) values further demonstrate the accuracy of predictions. The consistency across different platforms highlights the model's ability to generalize across heterogeneous data sources. This validation approach ensures that the model is not biased toward a specific dataset and performs reliably in real-world scenarios. The strong correlations observed in Fig. 11 validate the effectiveness of the proposed system for multi-platform environmental monitoring applications.

Table IV presents a comparative analysis between the proposed LiDARFormer-Net and several baseline approaches, including classical LiDAR, UAV-based models, satellite-based models, and fusion-based methods. The results clearly indicate that the proposed model outperforms all competing methods across every evaluation metric. Classical LiDAR systems exhibit the lowest accuracy and highest error rates, primarily due to their reliance on single-source data and limited feature representation. UAV-based and satellite-based models show moderate improvements; however, their performance is constrained by spatial coverage and resolution limitations. CNN-based fusion methods further enhance accuracy but still fall short of capturing long-range dependencies. Transformer-based fusion approaches demonstrate significant

improvements, highlighting the importance of attention mechanisms. Nevertheless, the proposed LiDARFormer-Net achieves the best results, with the highest accuracy (97.32%), lowest RMSE (0.053), and highest R^2 (0.983). These improvements are attributed to the integration of transformer-based learning and multi-platform data fusion, which enables comprehensive feature representation and robust prediction. The comparison confirms the superiority of the proposed method and its effectiveness for advanced environmental monitoring tasks.

TABLE IV. COMPARISON WITH THE STATE-OF-THE-ART STUDIES

Method	Data Source	Fusion Type	RMSE ↓	R^2 ↑	Accuracy (%) ↑	Real-Time
Classical LiDAR [39]	Single	No	0.312	0.612	72.35	No
UAV-based Model [40]	UAV	No	0.232	0.721	80.47	Yes
Satellite-based Model [41]	Satellite	No	0.198	0.783	83.26	Yes
CNN-based Fusion [42]	Multi	Basic Fusion	0.128	0.891	89.14	Yes
Transformer-based Fusion [43]	Multi	Advanced Fusion	0.087	0.948	94.21	Yes
LiDARFormer-Net (Proposed)	Multi	AI-Based Fusion	**0.053**	**0.983**	**97.32**	Yes

V. DISCUSSION

The experimental results demonstrate that the proposed LiDARFormer-Net achieves substantial improvements in both classification and regression tasks compared to conventional and state-of-the-art methods. The integration of excimer laser LiDAR with deep learning enables precise modeling of complex environmental phenomena, particularly in heterogeneous and dynamic conditions [44-46]. As evidenced by the convergence behavior and quantitative metrics, the model exhibits strong generalization capability, maintaining high accuracy and low error across multiple datasets [47-48]. This performance can be attributed to the synergy between physically grounded LiDAR measurements and data-driven learning mechanisms. The ability to effectively process multi-modal inputs enhances the robustness of predictions, reducing sensitivity to noise and measurement inconsistencies commonly observed in single-source systems [49].

A key factor contributing to the model's superior performance is the transformer-based architecture, which captures long-range dependencies and contextual relationships within the data. Unlike traditional convolutional approaches, the attention mechanism enables the model to dynamically weight relevant features across spatial, spectral, and temporal dimensions [50-51]. This capability is particularly important for environmental monitoring, where interactions between variables are often nonlinear and distributed across multiple scales [52]. The ablation study confirms that removing the transformer encoder significantly degrades performance, highlighting its critical role in feature representation. Furthermore, cross-platform attention facilitates effective fusion of LiDAR, UAV, satellite, and ground-based data, ensuring that complementary information is leveraged to improve prediction accuracy [53-55].

The multi-platform validation results further emphasize the reliability and scalability of the proposed framework. Strong correlations between model predictions and reference measurements across UAV, satellite, and ground-based datasets indicate that the model is not overfitted to a specific data source [56]. Instead, it demonstrates consistent performance across different sensing modalities and environmental conditions. This generalization capability is essential for real-world deployment, where data heterogeneity and variability are unavoidable [57]. Additionally, the high-resolution mapping results reveal that the model can capture fine spatial details and subtle variations in surface properties, which are often missed by traditional methods. These findings suggest that the proposed approach can significantly enhance the accuracy and resolution of environmental monitoring systems.

Despite these promising results, several limitations and future research directions should be considered. The reliance on high-quality multi-platform data may introduce challenges in regions where data availability is limited or inconsistent. Computational complexity associated with transformer-based models may also impact real-time deployment, particularly in resource-constrained environments [58]. Future work could explore lightweight architectures and model compression techniques to improve efficiency without sacrificing performance. Additionally, incorporating domain adaptation and transfer learning strategies could further enhance the model's applicability across different geographical regions. Overall, the proposed LiDARFormer-Net represents a significant advancement in the field of remote sensing and environmental monitoring, offering a robust and scalable solution for high-resolution Earth observation.

VI. CONCLUSION

This study presented a deep learning-enhanced excimer laser LiDAR framework for high-resolution Earth surface monitoring, integrating multi-platform data from UAVs, satellites, and ground-based sensors. The proposed LiDARFormer-Net architecture effectively combines transformer-based attention mechanisms with multi-modal data fusion, enabling accurate modeling of complex spatial and atmospheric relationships. Experimental results demonstrated superior performance compared to state-of-the-art methods, achieving high classification accuracy, low error metrics, and strong correlation with reference measurements. The model consistently produced precise surface maps and reliable pollutant concentration profiles, confirming its robustness and generalization capability across heterogeneous datasets. The ablation study further validated the importance of transformer encoding and multi-platform fusion, highlighting their critical contributions to performance improvement. Additionally,

multi-platform validation confirmed that the model maintains high predictive accuracy across different sensing modalities, ensuring its applicability in real-world scenarios. Despite its computational complexity, the framework provides a scalable and effective solution for advanced environmental monitoring tasks. Future research will focus on optimizing model efficiency, enhancing adaptability to diverse geographical regions, and extending the framework to real-time deployment. Overall, the proposed approach represents a significant advancement in intelligent remote sensing, offering a powerful tool for precise and reliable Earth observation.

REFERENCES

[1] Tafone, D., McEvoy, L., Meng Sua, Y., Rehain, P., & Huang, Y. (2023). Surface material recognition through machine learning using time of flight lidar. Optics Continuum, 2(8), 1813-1824.

[2] Yin, S., Sun, F., Liu, W., Bi, Z., Liu, Q., & Tian, Z. (2023). Remote identification of oil films on water via laser-induced fluorescence LiDAR. IEEE Sensors Journal, 23(12), 13671-13679.

[3] Kutova, O., Bidnyk, S., Yadav, K., Balakrishnan, A., Albert, J., & Jean-Ruel, H. (2025, May). Scalable phase trimming of non-hydrogen-loaded planar lightwave circuits with a krypton fluoride laser. In 2025 Photonics North (PN) (pp. 1-2). IEEE.

[4] Jegal, N. J., Periyasamy, M., & Balasubramanian, S. (2024, March). Technical review on sustainable low cost laser based weeder for autonomous and non-autonomous agriculture offroad vehicles. In AIP Conference Proceedings (Vol. 3035, No. 1, p. 020016). AIP Publishing LLC.

[5] Omarov, B., & Altayeva, A. (2018, January). Towards intelligent IoT smart city platform based on OneM2M guideline: smart grid case study. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 701-704). IEEE.

[6] Brown, C. E. (2025). Laser fluorosensors for oil spill detection. In Oil Spill Science and Technology (pp. 525-544). Elsevier.

[7] Dolgii, S. I., Makeev, A. P., Nevzorov, A. V., Nevzorov, A. A., Salnikova, N. S., & Kharchenko, O. V. (2025). Vertical Distribution of Ozone in the Upper Troposphere–Stratosphere according to lidar Sounding Data at the Siberian Lidar Station in 2023. Atmospheric and Oceanic Optics, 38(2), 161-165.

[8] Ikram, Z., & Omarov, B. (2024). Fusion of vision transformers and convolutional networks for advanced face anti-spoofing. In DTESI.

[9] Lyashenko, A. I., Volodina, E. M., Gol'din, Y. A., & Gureev, B. A. (2022). YAG: Nd3+ laser systems for marine UV Lidar. Journal of Communications Technology and Electronics, 67(12), 1475-1478.

[10] Grant-Jacob, J. A., Mills, B., & Zervas, M. N. (2023). Acoustic and plasma sensing of laser ablation via deep learning. Optics Express, 31(17), 28413-28422.

[11] Olzhayev, O., Imanbayeva, N., Mamikov, S., & Baibek, B. (2025). A Novel YOLO-Like Multi-Branch Architecture for Accurate Apple Detection and Segmentation Under Orchard Constraints. International Journal of Advanced Computer Science & Applications, 16(11).

[12] Liu, F., Lei, Y., Xie, Y., Li, X., Nan, Q., & Yue, L. (2023). Vehicle identification using deep learning for expressway monitoring based on ultra-weak FBG arrays. Optics express, 31(10), 16754-16769.

[13] Kutova, O., Bidnyk, S., Yadav, K., Balakrishnan, A., Albert, J., & Jean-Ruel, H. (2025). Scan-Free KrF Laser Phase Tuning of Non-Hydrogen Loaded Planar Lightwave Circuit Interferometer Arrays.

[14] Altayeva, A., Omarov, B., Suleimenov, Z., & Im Cho, Y. (2017, June). Application of multi-agent control systems in energy-efficient intelligent building. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS) (pp. 1-5). IEEE.

[15] Sharma, H., Pathak, R., Tanjum, S., Tonni, S. D., Treeno, J. T. S., & Al Noman, A. (2026). Impact of Laser in the Manufacturing Industry. Laser Materials Processing and Manufacturing Techniques, 105.

[16] Kulambayev, B. O., Olzhayev, O. M., Altayeva, A. B., & Zhunisbekova, Z. (2025). A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment. International Journal of Advanced Computer Science & Applications, 16(12).

[17] Romanovskii, O. A., Yakovlev, S. V., Sadovnikov, S. A., Nevzorov, A. A., Nevzorov, A. V., Kharchenko, O. V., ... & Kistenev, Y. V. (2025). Ground-based Stationary Differential Absorption Lidars for Monitoring Greenhouse Gases in the Atmosphere. Atmospheric and Oceanic Optics, 38(3), 345-359.

[18] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[19] Li, H., Andreev, M. V., Panchenko, Y. N., & Puchikin, A. V. (2022). Improving the stability of the optical system of a laser source based on a position-sensitive detector. Atmospheric and Oceanic Optics, 35(5), 615-619.

[20] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[21] Zhou, H., Gao, B., & Wu, W. (2023). Automatic crack detection and quantification for tunnel lining surface from 3D terrestrial LiDAR data. Journal of Engineering Research, 11(2), 239-257.

[22] Ding, J., Zhu, S., Xiao, Y., Zhou, Z., Siraj, M., Shi, Y., & Yu, Y. (2024). Randomized metalens array homogenizer for enhanced laser beam shaping. Optics Express, 32(23), 40514-40522.

[23] Altayeva, A., Omarov, B., & Im Cho, Y. (2017, December). Multi-objective optimization for smart building energy and comfort management as a case study of smart city platform. In 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 627-628). IEEE.

[24] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[25] Ibrayev, S., Omarov, B., Amanov, B., & Momynkulov, Z. (2024). Development of a deep learning-enhanced lower-limb exoskeleton using electromyography data for post-neurovascular rehabilitation. Engineered Science, 31, 1269.

[26] Rajput, V., & Bhardwaj, S. (2026). Laser machining of glass substrates. In Micro-Fluidic and Micro-electromechanical System Applications (pp. 202-227). CRC Press.

[27] Kulambayev, B., & Olzhayev, O. (2025). A Mask R-CNN Algorithm for Automated Segmentation of Asphalt Road Cracks. Procedia Computer Science, 269, 39-48.

[28] Tang, X., Liu, M., Zhang, Y., Liu, P., Shi, F., Qin, F., & Wang, Y. (2025). Thickness effect on the laser-induced thermoelectric voltage and laser-communication realization of the solution–derived inclined Bi2201 thin films. Journal of Sol-Gel Science and Technology, 115(3), 1757-1763.

[29] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[30] Jin, X., Fan, G., Zhang, T., Zhang, B., Mu, X., Xiang, Y., ... & Liu, W. (2025). Performance simulations for a spaceborne ozone lidar mission. Optics Express, 33(4), 6966-6986.

[31] Igor V. Malyk, Yevhen Kyrychenko, Mykola Gorbatenko, Taras Lukashiv, "Data Optimization through Compression Methods Using Information Technology", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.84-99, 2025. DOI:10.5815/ijitcs.2025.05.07

[32] Jia, L., Lin, H., Zhang, B., Cao, G., Chen, F., & Jia, B. (2025). Laser-nanofabrication-enabled multidimensional photonic integrated circuits. Photonics Insights, 4(2), R05-R05.

[33] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[34] Xia, J., Zhang, J., Jiao, H., Niu, X., Ji, X., He, T., ... & Wang, Z. (2026). Research progress and prospects of laser coating technology. Light: Advanced Manufacturing, 6(4), 870-895.

[35] Yang, Y., Lee, E., Park, Y., Seong, J., Kim, H., Kang, H., ... & Rho, J. (2025). The road to commercializing optical metasurfaces: current challenges and future directions. ACS nano, 19(3), 3008-3018.

[36] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[37] Xie, M., Jia, Y., Li, Y., Cai, X., & Cao, K. (2022). Experimental analysis on the optimal excitation wavelength for fine-grained identification of refined oil pollutants on water surface based on laser-induced fluorescence. Journal of Fluorescence, 32(1), 257-265.

[38] Sultan Mukhamedaly, Kymbat Kabekeyeva, Gulnar Mussabekova, Aliya Kuralbayeva, Bagdat Toibekova, Gulzhan Makashkulova, Batyrkhan Omarov, "A Pedagogical Framework for Ethical Skill Development in Higher Education within Smart Learning Environments", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 1-20, 2026. DOI:10.5815/ijmecs.2026.02.01

[39] Islam, M. N., Das, A. K., Billah, M. M., Rahman, K. S., Hiziroglu, S., Hattori, N., ... & Rudolfsson, M. (2023). Multifaceted laser applications for wood–a review from properties analysis to advanced products manufacturing. Lasers in Manufacturing and Materials Processing, 10(2), 225-250.

[40] Sun, D., Mao, J., Liu, M., Liu, H., Zhang, S., Li, B., ... & Ma, J. (2025). A fiber Bragg grating (FBG)-strain sensing tube for deep displacement measurement. Optics & Laser Technology, 188, 112938.

[41] Bobrovnikov, S. M., Gorlov, E. V., Zharkov, V. I., & Zaitsev, N. G. (2025). Laser Triggering System for Dual-pulse Laser Diagnostics. Atmospheric and Oceanic Optics, 38(4), 492-497.

[42] Olzhayev, O., Kulambayev, B., Sakenkyzy, N., & Belisbek, M. (2026). A Real-Time Multi-Scale Feature Pyramid YOLO Architecture for Accurate and Deployment-Efficient Road Damage Detection. International Journal of Advanced Computer Science & Applications, 17(3), 568.

[43] Saini, T. S., & Kalra, M. (2025). Emerging Trends and Future Directions in Aerospace Optical Sensing. Advanced Optical Sensors for Aerospace Applications, 393-406.

[44] Wang, G., Bai, Y., Li, Y., Yang, S., Zhang, X., Duan, W., & Lu, B. (2024). A high peak power and low peak-to-peak instability mid-infrared optical parametric oscillator pumped by a 1064 nm electro-optic cavity-dumped pulsed laser. Applied Physics B, 130(9), 161.

[45] Ding, N., Xi, Y., Jiang, W., Li, H., Su, J., Yang, S., & Lie, T. T. (2025). State-of-the-art carbon metering: Continuous emission monitoring systems for industrial applications. Heliyon, 11(3).

[46] Li, H., Wang, C., Cai, Y., Zhao, T., Gao, S., Yue, W., ... & Ding, F. (2026). Metasurface-empowered integrated silicon photonics: foundational principles, representative applications, and fabrication strategies. Advanced Photonics, 8(2), 024003-024003.

[47] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[48] Bobrovnikov, S. M., Gorlov, E. V., Zharkov, V. I., & Murashko, S. N. (2023). Estimation of energy and time parameters of laser radiation for efficient excitation of phosphorus oxide fluorescence. Atmospheric and Oceanic Optics, 36(5), 556-561.

[49] Ma, J., Kang, T., Ke, Z., Yao, M., Ma, X., Luo, Q., ... & Qin, J. (2024). Tunable Far-Infrared Polarization Imaging Based on VO2 Metasurfaces. Advanced Optical Materials, 12(11), 2302390.

[50] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[51] Kumar, R., Kumar, P., & Ranjan, A. (2025). Next Generation Optical Sensor Devices. In Advanced Optical Sensors: Noble Metal-Based Hybrid Composites (pp. 275-301). Singapore: Springer Nature Singapore.

[52] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[53] Mangesh P. Joshi, "Prioritization of Barriers to Digitization for Circular Systems using Analytical Hierarchy Process", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.3, pp.61-71, 2025. DOI:10.5815/ijitcs.2025.03.05

[54] Semenov, I. A., Rybaltovsky, A. A., Davydov, D. A., Likhachev, M. E., Lobanov, A. S., & Lipatov, D. S. (2025). Short Single-Frequency Fiber Lasers with Distributed Bragg Reflectors Based on Highly Yb-Doped Photosensitive Polarization-Maintaining Silica-Based Fiber. Bulletin of the Lebedev Physics Institute, 52(Suppl 11), S1104-S1120.

[55] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[56] Tahara, N., Nawa, S., Taira, R., Suyama, S., Hirotani, K., Maegami, Y., ... & Baba, T. (2024). Silicon photonics efficient fiber coupler based on a metastructure with a minimum feature size of 150 nm. Optics Express, 32(19), 34024-34033.

[57] Ikram, Z. (2025, May). Depth-Guided Neural Network for Robust Face Anti-Spoofing. In 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE.

[58] Nomoto, J., Koida, T., Yamaguchi, I., & Nakajima, T. (2024). Key sputtering parameters for precursor In2O3 films to achieve high carrier mobility. ACS Applied Materials & Interfaces, 16(46), 64113-64122.

Federated and Cross-Domain Student Performance Prediction

Sam Zhe Xuan[1], P. Ganesh Kumar[2], C. Rani[3], K. Kanagalakshmi[4]*, R. RajiniGanth[5], Atif Mahmood[6]*
Faculty of Data Science and Information Technology, INTI International University, Malaysia[1,6]
Department of Computer Science and Engineering-College of Engineering, Guindy,
Anna University, Chennai, Tamil Nadu, India[2]
Department of Computer Science and Engineering, Government College of Engineering,
Bodinayakanur, Tamil Nadu, India[3]
School of Science and Computer Studies, CMR University, Bangalore, Karnataka, India[4]
Department of Computer Science and Technology, SNS College of Engineering, Coimbatore, Tamil Nadu, India[5]

***Abstract*—Accurate student performance prediction is critical for data-driven educational decision-making; however, it is often hindered by data heterogeneity, privacy constraints, and domain shift across academic contexts. This study investigates student final grade (G3) prediction using three complementary machine learning paradigms: centralized learning, cross-domain generalization, and federated learning with personalization. Experiments were conducted on Portuguese and Mathematics student datasets using traditional regression models, ensemble methods, neural networks, and a personalized federated learning framework based on FedProx and FedBN. In the centralized setting, models capable of capturing non-linear relationships, particularly XGBoost and multi-layer perceptrons, achieved superior predictive performance, with XGBoost attaining an R^2 of 0.8308 and the lowest error metrics. In contrast, direct cross-domain application of models trained on Portuguese data to Mathematics outcomes resulted in severe performance degradation, with several models yielding negative R^2 values, highlighting the adverse impact of domain shift. To address privacy and heterogeneity challenges, a federated learning simulation was implemented. While the global federated model achieved moderate accuracy, the introduction of local personalization led to substantial performance gains. The personalized client models achieved stronger local predictive performance than the global federated model and showed competitive performance relative to centralized baselines. Learning-curve analysis further indicate that model performance in centralized settings improves with increasing data size but eventually plateaus, whereas cross-domain learning remains constrained despite additional data. In federated learning, predictive performance consistently improves across training rounds, demonstrating the effectiveness of iterative collaboration and client-level personalization. Overall, the results suggest that federated learning with personalization offers a competitive privacy-preserving alternative to centralized modeling and provides a clear improvement over direct cross-domain transfer in heterogeneous educational analytics.**

***Keywords*—Federated learning; cross-domain generalization; educational data mining; student performance; quality education**

I. Introduction

Educational institutions increasingly rely on data-driven approaches to understand, predict, and improve student academic performance [1]. Accurate early prediction of student outcomes enables timely interventions, optimized resource allocation, and informed pedagogical decision-making. With the growing availability of educational datasets, machine learning (ML) techniques have become central to learning analytics, offering superior predictive capabilities compared to traditional statistical methods [2]. However, deploying such models in real educational environments introduces challenges related to data heterogeneity, privacy preservation, and model interpretability [3], [4]. The prediction of student performance is inherently complex due to the multidimensional nature of educational data, which typically includes demographic, socio-economic, behavioral, and academic attributes [5].

Prior studies have extensively investigated a range of supervised learning models for predicting student academic performance, including linear regression variants, decision trees, ensemble methods, and neural networks, reporting varying levels of predictive accuracy [6]. Empirical evidence suggests that machine learning models can achieve very high performance, with some approaches attaining predictive accuracy of up to 97.12% [7]. In particular, tree-based ensemble techniques such as Random Forests, Gradient Boosting [8], and XGBoost [7] have consistently demonstrated superior performance due to their ability to capture complex nonlinear relationships and feature interactions [9]. However, despite their strong predictive capability, accuracy alone is insufficient in educational contexts, where transparency and explainability are essential to foster trust, support informed decision-making, and encourage adoption by educators and policymakers.

In addition to interpretability concerns, data privacy has emerged as a major barrier to the large-scale deployment of predictive analytics in education. Student data are highly sensitive and often governed by strict institutional and legal constraints, limiting the feasibility of centralized data collection across schools, campuses, or educational authorities. As a result, many existing studies assume centralized access to student records, which may not reflect realistic deployment scenarios. This gap motivates the exploration of privacy-preserving learning paradigms that can leverage distributed data while respecting institutional boundaries.

Federated learning (FL) is a decentralized learning framework in which multiple clients collaboratively train a shared global model without exchanging raw data. Instead, only

*Corresponding author

model updates, such as parameters or gradients, are communicated to a central server for aggregation, ensuring data privacy. This approach is well-suited to educational analytics, where institutions often exhibit heterogeneous (non-IID) data distributions due to differences in curricula, socio-economic factors, and assessment practices. However, standard FL methods such as Federated Averaging (FedAvg) often degrade under strong data heterogeneity, motivating the use of more robust variants such as Federated Proximal (FedProx) and Federated Batch Normalization (FedBN). Recent studies have also shown the feasibility of applying federated learning in privacy-preserving and resource-constrained distributed environments, including healthcare and network-based deployment scenarios [10].

In this study, heterogeneity refers to distributional differences between the Portuguese and Mathematics student datasets, including differences in feature distributions, target distributions, and learned model behavior across domains. Domain shift is operationally assessed through the degradation of R^2, MAE, and RMSE when a model trained on Portuguese data is evaluated on Mathematics data. Privacy is addressed at the simulation level by assuming that raw client data remain local during federated training, while only model parameters or updates are shared. Therefore, the privacy claim is limited to data locality and does not imply formal privacy guarantees such as differential privacy or cryptographic protection.

A longstanding challenge in educational machine learning is balancing predictive accuracy with interpretability. Advanced models such as deep neural networks [11] and ensemble methods, including XGBoost, often achieve strong predictive performance but operate as black-box systems, limiting transparency and interpretability [12], [13]. In contrast, simpler models such as Linear Regression and Elastic Net offer clear coefficient-based interpretability but may struggle to capture the complex non-linear relationships present in educational and behavioral data [14], [15]. Building on prior studies that emphasize interpretable educational prediction frameworks [16], [17], this work extends traditional regression-based approaches to cross-domain and federated learning paradigms. Beyond centralized evaluations of Elastic Net Regression, Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), and XGBoost, two additional experimental settings are introduced to enhance generalization, fairness, and privacy across heterogeneous educational domains.

Interpretability is considered as an important contextual requirement for educational prediction systems. However, the experimental focus of this study is on comparing centralized learning, cross-domain generalization, and personalized federated learning rather than conducting a separate explainability analysis.

The main contributions of this study are as follows:

- A comprehensive comparison of regression, kernel-based, neural, and ensemble models for student performance prediction.
- Introduction of cross-domain generalization and federated learning (FedProx + FedBN) frameworks to enhance scalability, fairness, and privacy under decentralized educational settings.
- Cross-Domain Learning (POR $\rightarrow$ MAT): This experiment evaluates model generalizability by training on one subject domain (Portuguese language) and testing on another (Mathematics) within the UCI Student Performance dataset [18]. It simulates realistic academic scenarios where predictive models must adapt to diverse cohorts without retraining from scratch.
- Federated Learning (FedProx + FedBN): To address data privacy and institutional heterogeneity, a hybrid federated learning framework was implemented. The model integrates *FedProx* regularization [19] to stabilize client updates under heterogeneous federated optimization and *FedBN* [20] to preserve client-specific batch-normalization statistics under non-IID feature distributions. This approach enables decentralized training across multiple clients while supporting local personalization and data locality.

II. Related Work

Student performance prediction has been a central topic in educational data mining (EDM), with extensive studies employing machine learning (ML) algorithms to model academic outcomes. Early foundational work by Cortez and Silva [16] applied decision trees, random forests, neural networks, and regression models to the UCI Student Performance dataset, showing that prior academic grades were among the strongest predictors of final performance. The UCI repository also identifies the Cortez and Silva study as the introductory paper for the Student Performance dataset, which contains Mathematics and Portuguese course records [18]. However, models that strongly depend on prior grades may have limited usefulness for early intervention when such grades are unavailable or when prediction is required across different academic domains.

Existing studies can be grouped into four broad categories. The first group focuses on centralized EDM, where student records are pooled and models are trained under a single-domain assumption. The second group emphasizes interpretable prediction using regression, decision trees, or post-hoc explanation tools. The third group investigates high-performing ensemble or neural models, but often assumes centralized access to student data. The fourth group explores privacy-preserving or federated learning, but with limited attention to subject-level domain shift and client-level personalization. This study is positioned at the intersection of these categories by comparing centralized learning, cross-domain generalization, and personalized federated learning under a unified evaluation protocol.

Recent literature emphasizes the importance of explainable and fair AI in academic prediction systems. Elastic Net Regression [14] is useful because it combines L1 and L2 regularization, supporting variable selection while handling multicollinearity. Support Vector Regression (SVR) [15] captures non-linear relationships through kernel functions, while tree-based ensemble algorithms such as XGBoost [21] are effective for structured tabular prediction tasks. In student-performance prediction, ensemble models have frequently shown strong predictive performance, especially when the data contain non-linear feature interactions [17], [4], [22]. However, these models are less transparent than linear models and often

require post-hoc interpretation methods such as SHAP [12] or LIME [13]. Since the present study does not include a separate SHAP experiment in the results section, explainability is treated here as a contextual motivation rather than a claimed experimental contribution.

Beyond centralized modeling, cross-domain generalization remains an important but less frequently tested problem in EDM. A model trained on one academic subject may not generalize well to another subject because the source and target domains can differ in feature distributions, target distributions, and predictor–outcome relationships. In this study, the Portuguese $\rightarrow$ Mathematics experiment is therefore treated as a cross-domain transfer setting under possible covariate shift and representation mismatch. This framing is important because poor cross-domain performance does not necessarily imply that a model is weak; rather, it may indicate that relationships learned from the source domain are not stable in the target domain.

Federated learning (FL) has emerged as a privacy-preserving learning paradigm in which multiple clients collaboratively train a model without centralizing raw data. FedProx was proposed to address statistical and systems heterogeneity in federated networks by adding a proximal term to stabilize local optimization [19]. FedBN was later introduced to handle non-IID feature distributions by keeping batch-normalization statistics local while sharing other model parameters [20]. These two methods are relevant to educational analytics because different institutions, departments, or subject groups may have heterogeneous student populations and restricted data-sharing policies. However, most existing educational prediction studies still focus on centralized settings, while fewer studies jointly evaluate centralized learning, cross-domain transfer, and personalized federated learning.

The literature, therefore, reveals three main gaps. First, many student-performance prediction studies evaluate models within a centralized single-domain setting, which may not reflect real deployment conditions across different subjects or institutions. Second, cross-domain validation is still limited, even though academic models may be applied to cohorts or subjects that differ from the training domain. Third, federated learning has not been sufficiently examined for personalized student-performance prediction under subject-level heterogeneity. This study addresses these gaps by evaluating centralized learning, Portuguese $\rightarrow$ Mathematics cross-domain transfer, and a personalized federated learning framework using consistent regression metrics.

A. Student Outcome Prediction Using Regression, Decision Trees, and Neural Networks

Regression-based models have long been used in EDM to quantify the effects of predictors such as study time, attendance, parental education, and previous grades [23], [16]. These models remain valuable because their coefficients provide relatively transparent evidence about predictor importance. However, linear models may struggle when relationships between variables are non-linear or highly interactive [24]. Decision-tree methods address some of these limitations by producing human-readable rules and modeling interactions between categorical and continuous variables [23], [25]. Neural networks, including multilayer perceptrons, can learn higher-order interactions when sufficient data are available, but their internal decision mechanisms are typically less transparent to educational stakeholders [26].

B. Ensemble Tree Methods and XGBoost

Ensemble tree algorithms, such as random forest and gradient boosting, are widely used for tabular prediction because they combine multiple weak learners to improve accuracy and robustness. XGBoost is a scalable gradient-boosting framework that has been widely adopted for structured prediction tasks [21]. In educational datasets, ensemble methods can capture non-linear relationships and feature interactions more effectively than simple linear models [22]. Nevertheless, their predictions may require post-hoc explanation methods when used in decision-support settings, particularly where teachers or administrators need to understand why a student is predicted to be at risk.

C. Interpretability and Ethical Considerations in Educational AI

Interpretability is important in educational applications because predictive models may influence interventions, resource allocation, or academic support decisions. Opaque models can raise concerns related to fairness, accountability, and trust [27]. Rudin [28] argues that high-stakes applications should prefer interpretable models where possible rather than relying only on post-hoc explanations of black-box models. In education, this issue is especially relevant because teachers and institutional decision-makers may be reluctant to use AI-based recommendations if the reasoning behind predictions is unclear [29]. Therefore, model performance should be interpreted alongside transparency, practical usability, and deployment constraints.

D. Feature Engineering and Representation Learning

Educational data often include heterogeneous demographic, behavioral, social, and academic variables. Feature engineering uses domain knowledge to transform these variables into meaningful predictors, such as interaction terms or normalized behavioral indicators. Regularized regression models such as Elastic Net can support this process by selecting relevant variables while controlling multicollinearity [14], [30]. This makes them useful as transparent baselines in educational analytics.

In contrast, representation-learning approaches such as ensemble models and neural networks learn complex patterns directly from data. Models such as XGBoost and MLPs can improve predictive performance by modeling non-linear interactions, but this often reduces direct interpretability [21], [22]. Explainable AI methods such as SHAP and LIME can partially bridge this gap by estimating feature contributions for complex models [12], [13]. However, because the present study focuses primarily on comparing learning paradigms rather than conducting a full XAI analysis, interpretability is discussed as a supporting consideration rather than as a separate experimental claim.

E. Federated Learning and Personalization

Federated learning is suitable for settings where data are distributed across institutions and cannot be centrally pooled. Standard FL can still suffer under non-IID data because client distributions may differ substantially. FedProx addresses this problem by regularizing local updates so that client models do not drift too far from the global model during local training [19]. FedBN addresses feature-shift non-IID data by preserving local batch-normalization statistics, which allows clients to maintain domain-specific normalization behavior [20]. In student-performance prediction, these mechanisms are relevant because different subjects, schools, or departments may produce different data distributions. Personalization further allows each client to adapt the global model to its local data distribution, which is important when a single global model cannot fully represent all clients.

F. Model Comparison, Trade-Offs, and Research Gap

Prior studies show a recurring trade-off between interpretable models, which are easier to justify but may have lower predictive capacity, and complex models, which may provide stronger prediction but lower transparency. However, fewer studies evaluate this trade-off together with cross-domain generalization and federated personalization. This gap is important because real educational analytics systems may face three simultaneous constraints: limited data sharing, distributional differences across domains, and the need for reliable local predictions.

This study, therefore, compares centralized, cross-domain, and federated learning settings using the same student-performance dataset family and the same regression metrics. The centralized experiment evaluates predictive performance when data pooling is possible. The cross-domain experiment tests whether a model trained on Portuguese records can generalize to Mathematics records. The federated experiment evaluates whether collaborative training with personalization can improve local adaptation while preserving data locality.

Table I summarizes the positioning of this study relative to prior work. Existing studies commonly focus on centralized accuracy, interpretability, cross-domain transfer, or privacy-preserving learning in isolation. In contrast, this work evaluates these issues together by comparing centralized learning, Portuguese $\rightarrow$ Mathematics cross-domain generalization, and personalized federated learning under a consistent experimental protocol. This structure directly addresses the need for educational prediction models that are accurate, privacy-aware, and robust to domain heterogeneity.

III. Methodology

The methodological framework of this study, illustrated in Fig. 1, was designed to systematically evaluate and compare machine learning approaches for predicting student academic performance. The framework is organized into three major experimental phases: **1)** centralized regression and ensemble learning baselines, **2)** cross-domain generalization across subject-specific datasets, and **3)** a privacy-preserving federated learning framework integrating *FedProx* and *FedBN*. All experiments were conducted using the UCI Student Performance dataset [18], which comprises demographic, behavioral, and academic attributes of secondary school students enrolled in Portuguese and Mathematics courses. Within this framework, a comparative modeling strategy was employed to predict students' final academic outcomes in Mathematics (`G3_mat`) and Portuguese (`G3_por`) using three representative machine learning paradigms: Elastic Net Regression, Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost) were selected to balance predictive complexity, interpretability, and robustness across heterogeneous educational datasets. Elastic Net offers transparent linear modeling with resilience to multicollinearity, making it suitable for interpretable academic performance analysis [14], [30]. MLP is employed to capture complex non-linear relationships among educational indicators [31], [32], while XGBoost delivers strong predictive capability through ensemble-based learning and feature interaction modeling [21]. Beyond centralized evaluation, these models are further integrated within a Federated learning framework, enabling privacy-preserving collaborative training across distributed institutional datasets without direct data sharing. This federated setup facilitates the assessment of global generalization, domain heterogeneity, and personalization trade-offs, thereby extending traditional educational data mining toward scalable, secure, and real-world multi-institutional analytics [23], [33].

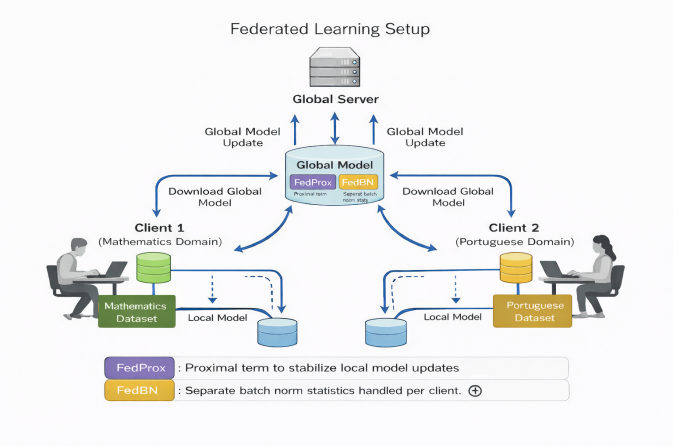


Fig. 1. Overall methodological framework for centralized learning, cross-domain generalization, and personalized federated learning.

A. Dataset Description

1) Data preprocessing and experimental setup: A unified, reproducible preprocessing framework was employed across three experimental settings to ensure fair comparisons across centralized, cross-domain, and federated learning paradigms.

a) Experiment 1 (Combined dataset): The Portuguese and Mathematics student datasets were merged using their shared feature space, excluding intermediate and final grade variables ($G1$, $G2$, and $G3$). A unified target variable, $G3_{\text{combined}}$, was constructed by averaging the final grades from both domains, and instances with missing target values were removed. The resulting dataset was separated into input features and target labels. A preprocessing pipeline based on a `ColumnTransformer` was applied, with numerical features imputed using mean substitution and categorical features transformed via one-hot encoding with unseen-category

TABLE I. COMPARISON OF PRIOR STUDIES ON STUDENT PERFORMANCE PREDICTION AND THE POSITIONING OF THIS WORK

Study Category	Learning Paradigm	Model Types	Key Limitations in Prior Work	Positioning of this Study
Early EDM Studies [16]	Centralized	Decision Trees, Random Forests, Neural Networks, Regression	Mostly single-domain evaluation; strong dependence on prior grades	Uses the same dataset family but evaluates centralized, cross-domain, and federated settings
Regression-Based EDM [14], [23]	Centralized	Linear Regression, Elastic Net	High interpretability but limited capacity for complex non-linear relationships	Uses interpretable regression models as baselines against non-linear models
Ensemble and Neural Models [21], [22]	Centralized	XGBoost, Random Forests, MLPs	Strong prediction but reduced transparency and limited deployment analysis	Compares high-capacity centralized models with cross-domain and federated alternatives
Explainable AI in Education [12], [13]	Post-hoc XAI	SHAP, LIME	Explanations often studied separately from privacy and domain-shift constraints	Uses explainability literature to motivate transparency concerns, without claiming a separate XAI experiment
Cross-Domain Learning in EDM	Centralized Transfer	Various ML Models	Generalization across subjects is often assumed rather than explicitly tested	Evaluates Portuguese → Mathematics transfer and identifies domain-shift effects
Federated Learning [19], [20]	Federated / Personalized	FedProx, FedBN, Personalized FL	Limited use in student-performance prediction; personalization often underexplored	Evaluates personalized FL for heterogeneous educational data while keeping raw data local
This Work	**Centralized, Cross-Domain, Federated**	**Elastic Net, SVR, XGBoost, MLP, FL Neural Models**	**Need for unified comparison under privacy and domain-shift constraints**	**Provides a unified evaluation showing strong centralized baselines, weak direct cross-domain transfer, and improved local adaptation through personalized FL**

handling. The processed data were subsequently split into training and test sets at a 70:30 ratio and integrated into end-to-end pipelines for Linear Regression, Support Vector Regression (SVR), Elastic Net, Multi-Layer Perceptron (MLP), and XGBoost models.

b) Experiment 2 (Cross-domain evaluation): To simulate a realistic source–target transfer learning scenario, the Portuguese and Mathematics datasets were treated as independent domains. Feature matrices and corresponding targets were extracted independently after excluding grade-related variables. In this setting, the preprocessing transformer was trained exclusively on the Portuguese dataset and applied to the Mathematics dataset without refitting. This design ensured that numerical imputation statistics and categorical encoding schemes were learned solely from the source domain, enabling an unbiased assessment of model generalization under domain shift. The resulting preprocessing pipeline was consistently used with XGBoost, SVR, Elastic Net, and MLP models.

c) Experiment 3 (Federated learning simulation): For federated learning, a global preprocessing strategy was adopted to guarantee feature consistency across decentralized clients while preserving data locality. Feature matrices from both domains were concatenated only for fitting a shared preprocessing transformer, without exposing labels or performing centralized model training. The federated preprocessor applied mean imputation followed by standardization for numerical attributes, which is essential for stable neural network optimization, and one-hot encoding for categorical features. The fitted transformer was then applied independently to each client dataset, producing locally encoded feature representations with identical dimensionality. These encoded inputs defined the neural network input space and were subsequently used within a federated optimization framework that employed FedProx with Batch Normalization personalization (FedBN), enabling privacy-preserving collaborative learning across heterogeneous educational data distributions.

B. Evaluation Metrics

All experiments were evaluated using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to ensure consistent and fair comparison across centralized, cross-domain, and federated learning setups. The R^2 metric measures the proportion of variance in the target variable explained by the model and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

where, y_i denotes the true target value, $\hat{y}_i$ is the corresponding model prediction, and $\bar{y}$ represents the mean of the observed targets. The Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors and is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

while the Root Mean Squared Error (RMSE) emphasizes larger errors through quadratic penalization and is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3}$$

In regression-based evaluation, no single performance metric is universally optimal, as each metric captures different aspects of model behavior. Widely used measures such as the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) provide complementary perspectives on predictive accuracy and error distribution. Recent studies highlight important nuances in metric selection. Chicco et al. [34] argue that R^2 is often more

informative than error-based metrics, as it offers a holistic assessment of model fit by quantifying the proportion of variance explained, whereas MAE and RMSE may exhibit limitations in interpretability when used independently. Conversely, Plevris et al. [35] emphasize the practical strengths of error-based measures, noting that MAE is robust to noisy datasets and less sensitive to outliers, while RMSE provides a balance between robustness and precision by assigning higher penalties to larger errors. Additionally, mean squared error (MSE) is particularly effective when minimizing the impact of substantial deviations is a primary objective. More recently, Dumre et al. [36] stress that evaluation metric selection should be aligned with specific research objectives and dataset characteristics, as different applications prioritize overall model fit, outlier sensitivity, or predictive precision. Accordingly, employing a multi-metric evaluation strategy is recommended to ensure a comprehensive and reliable assessment of regression model performance.

For federated learning experiments, performance was evaluated at both the global and local levels. *Global model performance* measures the predictive capability of the aggregated federated model and reflects the effectiveness of collaborative knowledge sharing across clients. In contrast, *personalized local performance* evaluates each client's locally adapted model on its own data distribution, capturing domain-specific learning and personalization effects. Reporting both metrics enables a comprehensive analysis of the trade-off between global generalization and client-level adaptation in heterogeneous federated environments.

IV. Results and Discussion

A. Overview

This section presents and analyzes the results from three experimental phases:

- Centralized regression and ensemble baselines,
- Cross-domain generalization, and
- Federated learning (FedProx + FedBN).

Performance was evaluated using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The objective was to assess not only the predictive accuracy of the models but also their interpretability, robustness, and scalability across data distributions.

This study investigates multiple machine learning paradigms for predicting student academic performance, including centralized learning, cross-domain generalization, and federated learning with personalization. The primary objective across all experiments is to predict final student grades ($G3$) using Portuguese and Mathematics student datasets under varying data availability and distributional assumptions.

B. Experiment 1: Centralized Learning on Combined Dataset

In the first experiment, all student records from the Portuguese and Mathematics datasets were centrally aggregated, and several regression models were evaluated, including Linear Regression, XGBoost, Support Vector Regression (SVR), Elastic Net, and a Multi-Layer Perceptron (MLP) Regressor (see Table II).

Among all evaluated models, XGBoost achieved the strongest predictive performance, attaining an R^2 score of 0.8308, along with a Mean Absolute Error (MAE) of 0.9356 and a Root Mean Squared Error (RMSE) of 1.5558. This superior performance highlights the effectiveness of tree-based ensemble methods in capturing complex, non-linear relationships present in student performance data when a rich, centrally available dataset is accessible.

The MLP Regressor also demonstrated competitive performance, achieving an R^2 of 0.7830, MAE of 1.1415, and RMSE of 1.7618. This result further suggests the presence of non-linear dependencies among educational features that neural networks are well suited to model.

In contrast, simpler models such as Linear Regression ($R^2 = 0.7681$), Elastic Net ($R^2 = 0.7441$), and SVR ($R^2 = 0.7229$) produced respectable, but comparatively lower predictive accuracy. While these models maintain interpretability and stability, their limited representational capacity restricts their ability to fully model the underlying complexity of student achievement patterns.

1) Conclusion for Experiment 1: When data can be centrally aggregated, models capable of learning non-linear feature interactions—particularly XGBoost and MLP—provide robust and highly accurate predictions, making them well suited for centralized student performance prediction tasks.

TABLE II. Performance of Centralized Models on the Combined Portuguese and Mathematics Student Dataset.

Model	Training / Evaluation	R^2	MAE	RMSE
Linear Regression	Combined (POR + MAT)	0.7681	1.1535	1.8212
XGBoost	Combined (POR + MAT)	**0.8308**	**0.9356**	**1.5558**
SVR	Combined (POR + MAT)	0.7229	1.1451	1.9908
Elastic Net	Combined (POR + MAT)	0.7441	1.1813	1.9131
MLP Regressor	Combined (POR + MAT)	0.7830	1.1415	1.7618

C. Experiment 2: Cross-Domain Generalization

The second experiment examined the feasibility of cross-domain generalization by training models exclusively on the Portuguese dataset and evaluating them on the Mathematics dataset. This scenario simulates real-world conditions where models trained in one educational context are applied to a different institutional or curricular setting (see Table III).

Across all evaluated models, performance deteriorated substantially. XGBoost, despite its strong centralized performance, achieved only a modest R^2 score of 0.1443, with MAE and RMSE values of 3.0015 and 4.2327, respectively. More notably, SVR ($R^2 = -0.0445$), Elastic Net ($R^2 = -0.0602$), and the MLP Regressor ($R^2 = -0.0701$) all yielded negative R^2 scores, indicating predictive performance worse than a naive mean-based baseline.

1) Conclusion for Experiment 2: The severe and consistent degradation in performance across all models confirms the presence of a significant domain shift between the Portuguese and Mathematics student datasets. These findings demonstrate that direct cross-domain deployment of machine learning models without adaptation is highly ineffective and leads to unreliable predictions.

TABLE III. CROSS-DOMAIN PERFORMANCE OF MODELS TRAINED ON THE PORTUGUESE DATA AND TESTED ON MATHEMATICS DATA.

Model	R^2	MAE	RMSE
XGBoost	0.1443	3.0015	4.2327
SVR	-0.0445	3.3561	4.6764
Elastic Net	-0.0602	3.4408	4.7113
MLP Regressor	-0.0701	3.2265	4.7333

D. Experiment 3: Federated Learning with Personalization

The third experiment explored a federated learning (FL) framework combining FedProx regularization with FedBN-style local batch-normalization personalization, followed by local model personalization (see Table IV).

The globally aggregated federated model achieved a moderate R^2 score of 0.6694, with MAE of 1.5508 and RMSE of 2.2211 on the combined evaluation set. While this performance surpasses direct cross-domain application, it remains lower than the best centralized model, indicating that a single global federated model may not fully represent heterogeneous local data distributions.

In contrast, the personalized federated models exhibited substantial performance gains. The personalized Portuguese client achieved an R^2 score of 0.8564, MAE of 0.9286, and RMSE of 1.2233, while the personalized Mathematics client achieved an R^2 score of 0.7805, MAE of 1.6149, and RMSE of 2.1437. These results indicate that local personalization substantially improves client-specific predictive performance compared with the global federated model. However, direct comparison with centralized learning should be interpreted cautiously because the evaluation scopes differ [see Fig. (2) to Fig. (4)].

1) Conclusion for Experiment 3: Federated learning combined with local personalization proves highly effective for distributed and heterogeneous educational data. The global model provides a strong initialization, while local fine-tuning enables clients to adapt to their unique data distributions, resulting in improved client-specific performance compared with the global federated model.

TABLE IV. PERFORMANCE OF FEDERATED LEARNING MODELS USING FEDPROX AND FEDBN WITH GLOBAL AND PERSONALIZED EVALUATION.

Model Variant	Evaluation Scope	R^2	MAE	RMSE
Global FL Model (FedProx + FedBN)	Combined (Global)	0.6694	1.5508	2.2211
Personalized FL Model	Portuguese Client	**0.8564**	**0.9286**	**1.2233**
Personalized FL Model	Mathematics Client	0.7805	1.6149	2.1437

V. DISCUSSION AND CONCLUSION

This study provides several important insights into student performance prediction under centralized, cross-domain, and federated learning paradigms. First, the results demonstrate that when centralized data pooling is feasible, models with higher representational capacity—particularly ensemble and neural approaches—achieve superior predictive accuracy compared to linear and regularized regression models. This highlights the importance of modeling non-linear interactions among socio-academic features in educational datasets.

Second, the consistently poor performance observed across domains confirms that domain shift between academic subjects is a fundamental challenge. Models trained on Portuguese data fail to generalize effectively to Mathematics outcomes, even when trained with increasing amounts of source-domain data. These findings indicate that naïve cross-domain deployment without adaptation leads to unreliable predictions and is unsuitable for real-world educational settings. Most importantly, the federated learning experiments demonstrate that personalization is critical for learning in heterogeneous, privacy-constrained environments. While the global federated model captures shared patterns across institutions, client-level personalization enables effective adaptation to local data distributions. Personalized federated models outperform the global federated model and achieve competitive local performance, illustrating the value of combining shared global learning with client-specific adaptation. Overall, these findings suggest that federated learning with personalization offers a promising privacy-preserving alternative to centralized modeling, particularly when student data are distributed across heterogeneous academic domains.

A. Learning Curve Analysis

Learning curve analysis was used to further examine convergence behavior and generalization trends across the evaluated learning paradigms. In centralized learning, model performance stabilizes as training data increases, with non-linear models benefiting most from additional data, indicating effective convergence under single-domain conditions.

In contrast, cross-domain learning exhibits persistent generalization gaps, with test performance remaining low or even negative despite increased source-domain data. This behavior confirms that domain mismatch, rather than data scarcity, is the primary cause of performance degradation in cross-domain settings.

Federated learning shows consistent improvement across communication rounds, demonstrating effective collaborative optimization. Personalized models converge faster and achieve higher performance than the global federated model, reinforcing the importance of heterogeneity-aware learning strategies in decentralized environments.

B. Key Implications

The experimental results lead to three central implications:

- Centralized learning remains effective when data sharing is possible but is limited by privacy constraints and domain variability.
- Direct cross-domain transfer across academic subjects is ineffective without explicit adaptation mechanisms.
- Federated learning with client-level personalization provides a practical and high-performing solution for multi-institutional educational analytics.

Overall, the findings show that centralized learning is effective when data pooling is possible, but its practical use may be limited by institutional privacy constraints. Direct cross-domain transfer from Portuguese to Mathematics produced weak generalization, indicating that subject-level domain shift

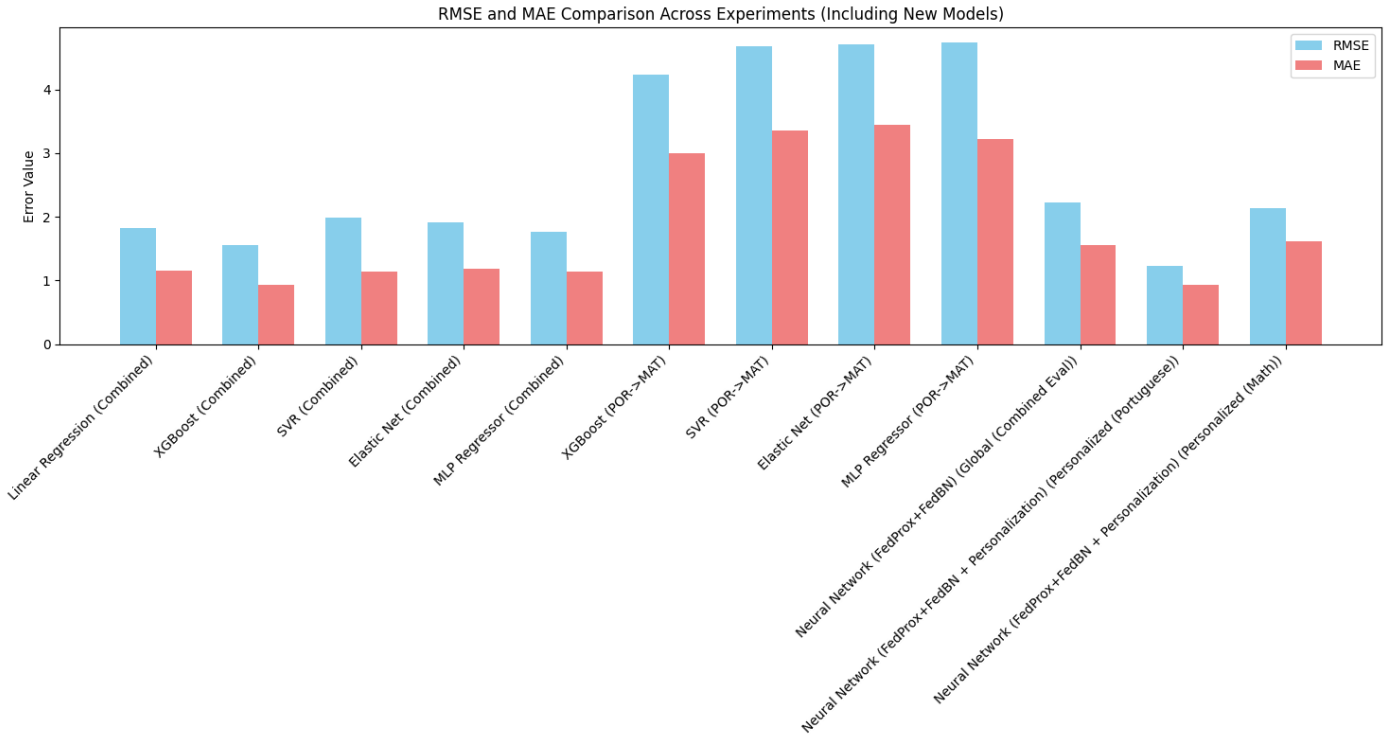


Fig. 2. RMSE and MAE comparison across centralized, cross-domain, and federated learning models.

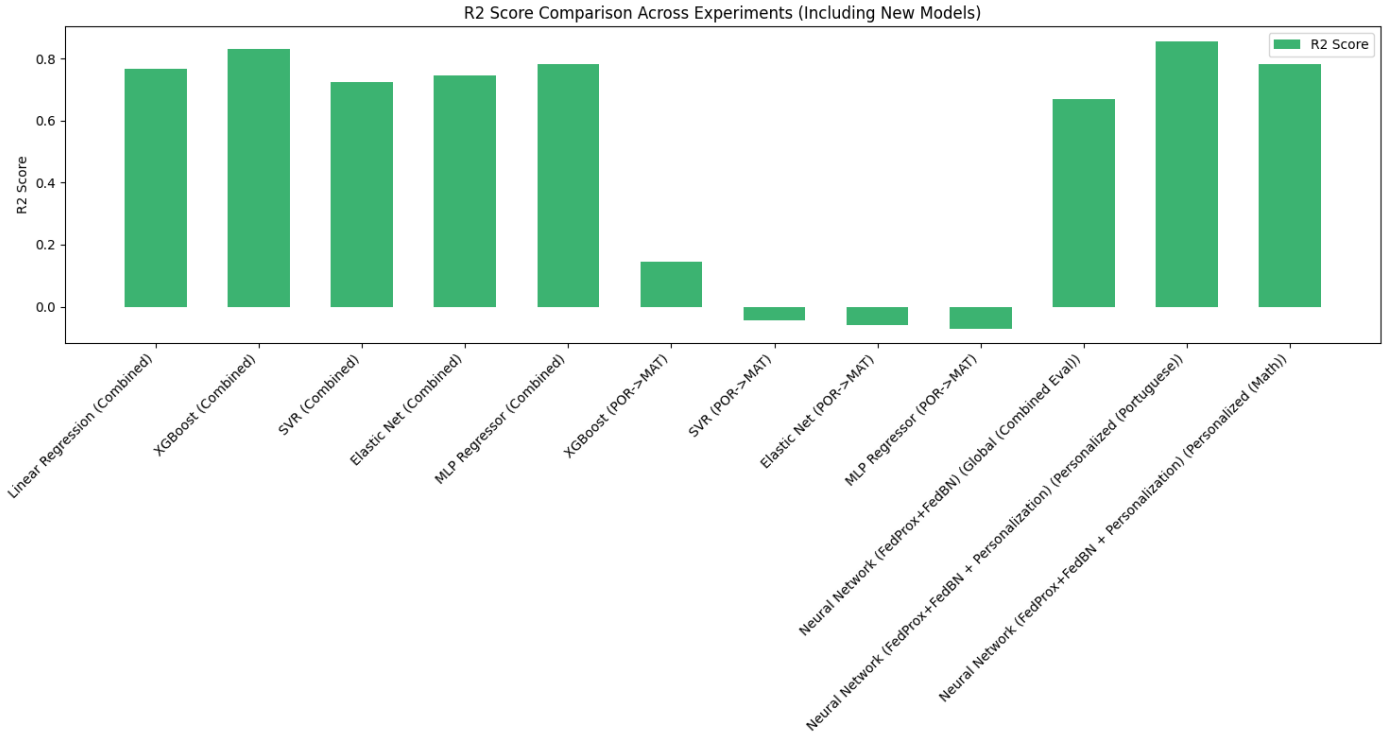


Fig. 3. R^2score comparison across centralized, cross-domain, and personalized federated learning models.

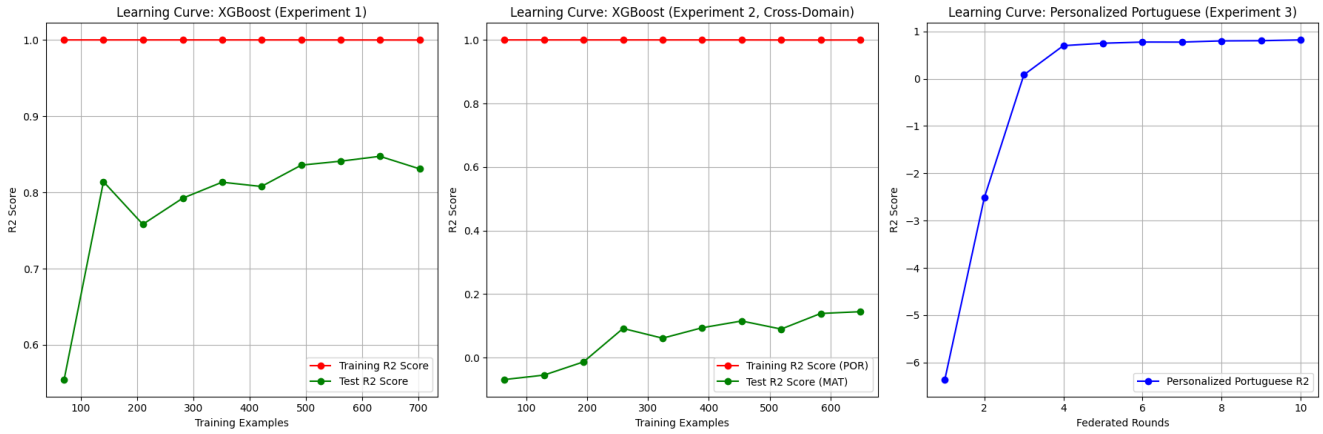


Fig. 4. Learning-curve comparison for centralized XGBoost, cross-domain transfer, and personalized federated learning.

must be explicitly addressed. The federated learning results show that a global model can capture shared patterns across domains, while client-level personalization improves local predictive performance. However, these findings are limited to the UCI Student Performance dataset and a simulated two-client federated setting. Future work should evaluate larger multi-institutional datasets, repeat experiments across multiple random seeds, report statistical confidence intervals, and measure communication cost under realistic federated deployment conditions.

REFERENCES

[1] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," *Kybernetes*, vol. 51, no. 2, pp. 916–934, 2022.

[2] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Scientific Reports*, vol. 15, no. 1, p. 26879, 2025.

[3] Y. Zhang, Y. Zhang, and J. Zhang, "Educational data mining techniques for student performance prediction: A review," *Frontiers in Psychology*, vol. 12, 2021.

[4] I. Issah, M. Asante, and S. Amankwah, "A systematic review of machine learning techniques for student performance prediction," *Heliyon*, vol. 9, no. 5, 2023.

[5] X. Bai, F. Zhang, J. Li, T. Guo, A. Aziz, A. Jin, and F. Xia, "Educational big data: Predictions, applications and challenges," *Big Data Research*, vol. 26, p. 100270, 2021.

[6] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student'performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[7] O. P. Ojajuni, F. Ayeni, O. Akodu, F. Ekanoye, S. Adewole, T. Ayo, S. Misra, and V. W. A. Mbarika, "Predicting student academic performance using machine learning," in *Communication Systems and Applications*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237550930

[8] A. Razaque and A. Alajlan, "Supervised machine learning model-based approach for performance prediction of students," *Journal of Computer Science*, vol. 16, no. 8, pp. 1150–1162, 2020.

[9] M. Arifin, W. Widowati, F. Farikhin, and G. Gudnanto, "A regression model and a combination of academic and non-academic features to predict student academic performance," *TEM Journal*, vol. 12, no. 2, p. 855, 2023.

[10] A. Mahmood, Z. H. Azizul, M. Zakariah, S. B. Belhaouari, A. Altameem, R. Ramli, A. S. Almazyad, M. L. M. Kiah, and S. R. Azzuhri, "Implementing federated learning over vpn-based wireless backhaul networks for healthcare systems," *PeerJ Computer Science*, vol. 10, p. e2422, 2024.

[11] D. Subramanian, A. Ajitha, and S. S. Maidin, "Unveiling hybrid model with naive bayes, deep learning, logistic regression for predicting customer churn and boost retention," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 1379–1391, 2025.

[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[16] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," https://www.researchgate.net/publication/228780408, 2008.

[17] M. Rashid, F. A. Khan, A. Alqahtani, and M. Hameed, "Classifying and predicting students' performance using popular algorithms," *Journal of Computer Science*, vol. 15, no. 8, pp. 1179–1190, 2019.

[18] UCI Machine Learning Repository, "Student performance dataset," 2014, introductory paper: Cortez and Silva, 2008.

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.

[20] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," in *International Conference on Learning Representations*, 2021.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[22] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and shapley additive explanations," *IEEE Access*, vol. 9, pp. 152 688–152 703, 2021.

[23] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 2010.

[24] M. A. Babyak, "What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models," *Psychosomatic Medicine*, 2004.

[25] M. Hoq, P. Brusilovsky, and B. Akram, "Analysis of an explainable student performance prediction model in an introductory programming course," 2023.

[26] S. G. Mundhe, S. Y. Gaikwad, and A. Professor, "Performance prediction in educational data mining using neural network," *International Journal of Innovative Research in Advanced Engineering*, vol. 8, pp. 582–585, 2019.

[27] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing*, 2024.

[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019.

[29] M. Lucas, Y. Zhang, P. Bem-haja, and P. N. Vicente, "The interplay between teachers' trust in artificial intelligence and digital competence," *Education and Information Technologies*, vol. 29, no. 17, pp. 22 991–23 010, 2024.

[30] J. K. Tay, N. Aghaeepour, T. Hastie, and R. Tibshirani, "Feature-weighted elastic net: Using features of features for better prediction," *Statistica Sinica*, vol. 33, no. 1, pp. 259–279, 2023.

[31] Y. Liu, C. Zhao, and Y. Huang, "A combined model for multivariate time series forecasting based on mlp-feedforward attention-lstm," *IEEE Access*, vol. 10, pp. 88 644–88 654, 2022.

[32] C. Enăchescu, "Approximation capabilities of neural networks," *Journal of Numerical Analysis, Industrial and Applied Mathematics*, vol. 3, no. 4, pp. 221–230, 2008.

[33] E. Kalita *et al.*, "Educational data mining: a 10-year review," *Information Retrieval Journal*, 2025.

[34] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.

[35] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," 2022.

[36] P. Dumre, S. Bhattarai, and H. K. Shashikala, "Optimizing linear regression models: A comparative study of error metrics," *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 1856–1861, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275789357

Intelligent Traffic Surveillance: Machine Learning-Based Detection of Vehicle Speed Violations

Niloy Kanti Paul, Dipanwita Saha, Kaushik Biswas, Sultanul Arifeen Hamim, Tanvir Ahmed, Rifath Mahmud
Department of Computer Science, American International University-Bangladesh, Dhaka-1229, Bangladesh

***Abstract*—With the rapid increase in the number of vehicles on roads, traffic management, and safety enforcement have become significant challenges worldwide. Traditional speed violation detection systems either employ high-end hardware, expensive computational resources, or post-processed video data, which are inefficient to implement in real time. This study presents a real-time intelligent vehicle speed violation detection system using YOLOv8 for object detection and SORT for vehicle tracking, and a new Speed Detection Algorithm (SDA). The system can effectively detect vehicles and calculate their speed from video recorded by low-cost fixed cameras. Unlike other models that process simulated or post-processed video data, the new model processes real-life scenarios such as changing lighting and weather conditions. Experimental results indicate that the system achieves 92% to 95% vehicle detection accuracy while maintaining a Mean Absolute Error (MAE) of 1.8 km/h and Root Mean Square Error (RMSE) of 2.5 km/h for speed estimation, and 98% effective at speed detection compared to various other systems that came before it in terms of real-time processing effectiveness. This cost-effective and scalable solution can be incorporated into traffic observation systems for the improvement of road safety and regulation of speed limit compliance.**

***Keywords*—Machine learning; speed detection; object detection; vehicle tracking; violation**

I. Introduction

The rapid technological advancements and urbanization in recent decades have led to a tremendous surge in the number of vehicles traveling along roads worldwide. As transport infrastructure continues to advance, more and more vehicles have come with them, accompanied by terrible problems in traffic flow, traffic jams, and road safety. Speed violation has been one of the worst among such problems, and a high rate of road accidents and casualties has followed from it [1]. Traditional speed detection has been mainly reliant on sensor and radar based technologies, which, with high performance, have high operation and maintenance fees and do not scale. Manual traffic enforcement, on the contrary, is inefficient and labor intensive, especially in high-density urban and rural areas with insufficient surveillance infrastructure. To bridge these gaps, computer vision and AI-based technologies have come with increasing interest in traffic monitoring and police enforcement. These technologies offer real-time and automatic detection and classification of vehicles, thereby increasing efficiency, reducing human interventions, and increasing road safety enforcement [2].

In recent years, various AI-based methods have been explored to estimate vehicle speed, e.g., Optical Flow Analysis, Support Vector Machines (SVMs), and Convolutional Neural Networks (CNNs) [3]. These methods have worked very effectively in moving object detection and tracking; however, there are various limitations. Most such methods require high-performance hardware and thus are not feasible to be implemented in low-end environments. Some methods rely upon synthetic or preprocessed data, which do not depict real-world traffic situations effectively and lead to inconsistencies when implemented in real world settings. Another limitation results from inefficiency in some methods to work in various environmental settings, e.g., in nighttime, rainy, or foggy environments when visibility reduces [4], [5]. To overcome these constraints, in this study, a smart traffic monitoring system using deep learning-based vehicle detection and tracking algorithms to detect vehicles and estimate velocity with high accuracy has been proposed. The system employs You Only Look Once version 8 (YOLOv8) for detection and Simple Online and Real-time Tracking (SORT) for tracking and a Speed Detection Algorithm (SDA) developed and implemented in-house to estimate velocity. Unlike conventional techniques with high-end hardware demands, our system can be implemented in low-cost hardware and, therefore, can be deployed in various traffic environments. Our proposed model has also been tested in various lighting and environmental conditions to provide robust performance in real-world environments. The primary concern in our research is to develop a computationally efficient, real-time, and scalable speeding violation detection system that can be implemented in already deployed traffic monitoring infrastructures with minimal computational overhead. The system employs machine learning-based detection and tracking algorithms to achieve high detection performance with real-time processing efficiency. Experimental results indicate that the proposed system achieves vehicle detection accuracy ranging from 92%–95%, while maintaining low speed estimation errors with an MAE of 1.8 km/h and RMSE of 2.5 km/h.

Key contributions of this work include:

- Development of a real-time, low-cost vehicle speed detection system using YOLOv8 for object detection and SORT for vehicle tracking.
- Design and implementation of a custom Speed Detection Algorithm (SDA) that accurately measures vehicle speed using video footage from low-cost cameras.
- Validation of the proposed model under diverse environmental conditions, including daylight, night-time, rainy, and foggy weather, ensuring robust real-world performance.

- Comparison with state-of-the-art approaches, demonstrating higher accuracy and lower processing time.
- Provision of a scalable solution that can be integrated into existing traffic management systems, contributing to enhanced road safety and effective law enforcement.

The remainder of the study is structured as follows: Section II presents a comprehensive literature review of existing speed detection systems. Section III details the proposed methodology, including dataset creation, model architecture, and speed calculation techniques. Section IV presents experimental results and comparative analysis, followed by discussions on limitations and future work in Section V. Finally, Section VI concludes the study.

II. Literature Review

Efficient traffic regulation has always been a problem in the majority of countries. With a rising number of vehicles traveling roads, traffic detection and speeding detection have always remained priority fields to research. Radar-based detection and man-based detection have always been adopted, which have been ineffective and expensive. This has prompted studies in machine learning and deep learning-based algorithms to automatically detect vehicles and estimate speeding. From conventional image processing to recent deep learning-based techniques have been proposed. This section provides a review of already existing studies in vehicle detection, tracking, and speeding estimation and classifies them into different methodological categories.

Early research in vehicle detection and estimation relied heavily upon frame-by-frame analysis-based techniques in image processing. One early attempt was by P.G. Michalopoulos, who suggested Autoscope, a real-time vehicle detection system that compared frames to determine motion. The system was equipped with a camera with a frame rate of 30 frames per second and a frame rate with a value of 97% accuracy. However, one key shortcoming with such a technique was that it could not be designed to be adaptable to varying weather and therefore was not robust in real-world environments [6]. Another conventional approach was to utilize motion-blurred images to estimate vehicle velocity. The technique, suggested by Lin et al., utilized Sum Modified Laplacian (SML) focus measure to determine the best frame and get estimates for speed. Though such a system had a frame rate with a value of 95% accuracy, it was impacted in terms of reliability in low-visibility environments and when cars had very high speeds [7]. On a related note, a Combination of Saturation and Value (CVS)-based technique was suggested by Karim and Dehghani that compared frames to determine motion. However, such a technique was computationally expensive and could analyze just 12 frames in a second, which had a very negative impact on real-time performance [8]. Though such conventional image processing-based techniques were useful in terms of learning about vehicle detection, they had high computational complexity, low dynamic environment adaptability, and low generalization to different lighting and weather environments.

The evolution of machine learning (ML) has significantly improved vehicle detection and speed estimation, allowing for more robust and adaptable systems. Pavlidis et al. introduced a Fuzzy Neural Network (FNN) classifier, which used infrared imaging signals to detect vehicles and count their numbers. However, despite using a dual-band camera system, the model's accuracy was lower than 50%, making it unsuitable for real-world deployment [9]. Another notable ML-based study was conducted by Li et al., who developed an Intelligent Visual Internet of Things (IVIoT)-based system for vehicle monitoring. Their approach integrated haze removal techniques to enhance image clarity under poor weather conditions. Despite its capability to extract vehicle labels from urban traffic footage, the model only achieved an accuracy rate of 85.80%, limiting its performance for large-scale applications [10]. For speed estimation, Shim et al. proposed an optical character recognition (OCR) based vehicle speed estimation method (VSEM). This method assessed speed by analyzing alterations in video playback speed and achieved an impressive 94.99% accuracy rate. However, its major drawback was its dependency on preprocessed or re-encoded video data, making real-time deployment infeasible [11]. These ML-based approaches introduced significant improvements over traditional image processing methods, particularly in object classification and adaptability. However, challenges such as low real-time efficiency, computational overhead, and dependence on high-quality datasets still need to be addressed.

Deep learning has revolutionized vehicle detection and estimation of vehicle speed with successful real-time traffic monitoring applications. Deep learning using Convolutional Neural Networks and advanced detection algorithms has been extensive in high-precision and scalable detection processes. One such pioneering contribution in this direction has been by Grent et al., wherein Region-Based CNN-based models have been applied in detection. It had a detection rate with a rate of accuracy as high as 78% with their approach, which utilized SORT tracking algorithm. It was, however, constrained by poor processing rates and poor real-time responsiveness [12]. A performance-optimized solution with a better use of DL was brought by Luvizon et al., wherein a vehicle speed estimation system in a video has been proposed using a Kanade–Lucas–Tomasi algorithm and T-HOG descriptors to detect license plates. Though it had a precision rate of 0.93 and recall rate of 0.87, it had a drastic limitation in detecting vehicle speeds at a distant location, as it required cameras to be installed very close to vehicles [13]. To enhance detection efficiency even more, Şentaş et al. combined a Histogram of Oriented Gradients (HOG), Support Vector Machines (SVM), and You Only Look Once (YOLO) for vehicle classification. Though it ensured better classification accuracy, it led to a problem with overfitting, reducing generalization across different sets [14]. Though huge breakthroughs were brought by deep learning, a lot of researches continued to rely upon high-end computational power, manipulated sets, or constrained environments, reducing real-world and low-budget deployment effectiveness.

A comparative analysis of prior work reveals intrinsic limitations in existing vehicle detection and speed estimation algorithms. One such limitation is the utilization of synthetic or pre-processed data and thus real-time applicability in question. V2I network-based and OCR-based speed estimation researches were very accurate but tested with simulations or re-encoded video and thus could not be performed in real-time traffic [15], [11], [16]. Most researches lack the flexibility to environmental and illumination factors and per-

form best in day-time and breaks in the night-time, fog, or rainy environments and thus reduces usability in real-time traffic [6], [10], [13]. Another limitation is utilizing high-end hardware and it becomes infeasible to be scaled in low-end environments. Most researchers utilize specialized cameras, IR sensors, or computationally heavy deep networks and thus prove infeasible to be mass-deployed, particularly in developing nations [7], [8], [12]. Furthermore, some researches focus on vehicle detection without real-time speed estimation and thus proves infeasible to deploy in automated policing [9], [17]. This research proposes a real-time, low-cost, and high-accuracy vehicle speed violation detection system trained on real-world data across multiple environmental conditions. The system efficiently runs in low-configuration hardware by utilizing a combination of YOLOv8 and SORT tracking to perform real-time, high-precision detection, tracking, and velocity estimation and thus provides a practical and scalable solution to traffic monitoring.

Table I presents a comparative overview of existing vehicle detection and speed estimation approaches. Earlier image-processing-based techniques suffered from limited adaptability to dynamic traffic conditions and environmental variations. Machine learning approaches improved classification performance but often lacked real-time capability or required computationally expensive hardware. Recent deep learning-based methods achieved higher detection accuracy; however, many systems remained dependent on constrained datasets or post-processed video inputs. In contrast, the proposed system combines YOLOv8, SORT, and a custom Speed Detection Algorithm to provide real-time vehicle monitoring with robust performance under diverse environmental conditions using low-cost surveillance infrastructure.

III. Methodology

To develop a robust and efficient vehicle speed violation detection system, a custom dataset was created from real-world traffic footage. The dataset was carefully designed to include vehicles under various environmental conditions, ensuring the model generalizes well in real-world scenarios. The dataset consists of 12,000 manually labeled images, extracted from video footage recorded on urban roads, highways, and intersections. The dataset includes six distinct vehicle categories: bikes, auto-rickshaws, cars, trucks, buses, and other vehicles.

A. Dataset Collection Process

Since publicly released dataset will be non-real-time applicable to speeding detection, a real-world video dataset has been set. The dataset generation was achieved in three steps in a principal manner: video collection, frame capture, and annotation to encompass traffic scenarios in a holistic fashion.

The first set consisted of high-resolution video frames from fixed roadside cameras installed along urban roads, highways, and intersections. The frames were recorded in various environmental conditions like day, night, fog, and rainy environments to provide real-world variability to the dataset. The video recording was recorded at a frame rate of 30 frames per second to provide a smooth tracking of vehicle motion and vehicle speed estimation. Following the video recording set, frame pulling was performed to pull frames from recorded video. Frames were pulled at regular time steps to provide temporal continuity to support correct tracking. A total of 12,000 frames were selected from different traffic scenarios to provide a representative set in terms of vehicle speeds, densities, and orientations. Frames were manually inspected to discard duplicate or low-quality frames to increase reliability in the dataset.

The final operation in dataset generation was annotation and labelling using LabelImg, which is a free and open-sourced labelling tool. Individual frames captured were manually labelled to generate bounding boxes around detected vehicles and respective class labels for object detection. Six categories of vehicles were present in the dataset: bikes, auto-rickshaws, cars, trucks, buses, and others. Timestamps for each frame were also captured to support vehicle tracking and estimation of vehicle speed precisely. Annotations were in YOLO format with each labelled object described by a Class ID, bounding box coordinates (x-center, y-center, width, height) and detection correctness in terms of a confidence score. Manual checking ensured high-quality labelling and removal of errors that would have otherwise affected model performance. This structured method to create a dataset ensured that a robust dataset representative, diversified, and suitable for real-world traffic environments was established, providing a solid basis to design a good vehicle speed detection system. Table II presents the classification of objects based on vehicle type.

Table III and Table IV summarize the distribution of vehicle categories and environmental conditions in the proposed dataset. The dataset was intentionally designed to include a diverse range of vehicle types and traffic scenarios to improve the model's generalization capability in real-world environments. Cars constituted the largest portion of the dataset due to their high frequency in urban traffic, while motorcycles, trucks, buses, and auto-rickshaws were also sufficiently represented. Additionally, the dataset incorporated challenging environmental conditions, including night-time, rainy, and foggy scenes, enabling the proposed system to learn robust vehicle representations under varying visibility and illumination conditions.

B. Data Annotation and Labelling

To ensure proper detection and tracking of vehicles, a manual annotation was carried out to mark vehicles in each frame extracted. The annotation was carried out using LabelImg, a widely used open-source tool for annotating images in deep learning. A bounding box and a class label were assigned to each vehicle in the dataset. The six types of vehicles considered in this study were bikes, auto-rickshaws, cars, trucks, buses, and other vehicles. Each frame was annotated carefully to give the precise location of vehicles to ensure bounding boxes correctly capture them. The annotation was in the YOLO format, in which each item annotated was represented in terms of a Class ID and normalized bounding box coordinates (x-center, y-center, width, height). This format best supports real-time detection of objects and effective model training. To maintain quality and consistency in annotation, different annotators worked simultaneously in annotation and cross-checked each annotation to minimize errors. Any erroneous or ambiguous bounding boxes were rectified to enhance detection. Furthermore, frame timestamps were recorded for each frame that had been annotated to facilitate vehicle tracking and

TABLE I. COMPARATIVE ANALYSIS OF EXISTING VEHICLE DETECTION AND SPEED ESTIMATION METHODS

Study	Methodology	Key Strength	Limitation	Real-Time Support
Michalopoulos [6]	Frame-based image processing	High detection accuracy	Poor adaptability to weather conditions	Partial
Lin et al. [7]	Motion blur analysis	Accurate speed estimation	Sensitive to visibility and motion blur	No
Karim and Dehghani [8]	CVS-based motion detection	Effective motion comparison	Computationally expensive	Limited
Pavlidis et al. [9]	Fuzzy Neural Network	Infrared-based detection	Low detection accuracy	No
Li et al. [10]	IVIoT-based monitoring	Weather enhancement capability	Moderate accuracy	Partial
Shim et al. [11]	OCR-based VSEM	High estimation accuracy	Requires post-processed video	No
Grents et al. [12]	CNN + SORT	Deep learning-based tracking	Low real-time efficiency	Limited
Luvizon et al. [13]	T-HOG + KLT tracking	Good precision and recall	Camera placement dependency	Partial
Şentaş et al. [14]	HOG + SVM + YOLO	Improved classification	Overfitting issues	Partial
Proposed System	YOLOv8 + SORT + SDA	Real-time, multi-weather, low-cost deployment	Minor errors at extreme speeds	Yes

TABLE II. CLASSIFICATION OF THE OBJECTS ACCORDING TO THE VEHICLE TYPE.

Classification Number	Class
0	Bike
1	Auto
2	Car
3	Truck
4	Bus
5	Other Vehicles

TABLE III. DATASET DISTRIBUTION ACROSS VEHICLE TYPES AND ENVIRONMENTAL CONDITIONS.

Category	Number of Images	Percentage (%)
Cars	4,200	35.0
Motorcycles / Bikes	2,300	19.2
Auto-Rickshaws	1,500	12.5
Trucks	1,700	14.2
Buses	1,200	10.0
Other Vehicles	1,100	9.1
Total	**12,000**	**100**

TABLE IV. ENVIRONMENTAL DISTRIBUTION OF THE DATASET

Condition	Number of Images	Percentage (%)
Daylight	5,400	45.0
Night-time	2,600	21.7
Rainy	2,100	17.5
Foggy	1,900	15.8
Total	**12,000**	**100**

velocity estimation. Such timestamps played a very important role in velocity estimation from frame positional differences.

The annotated dataset was then put to use as ground truth to test and train the model to make sure that vehicle detection algorithm would be robust and effective in detecting and classifying vehicles in different real-world settings. The labelling exercise proved useful in making the proposed vehicle speeding violation detection system robust and effective.

C. Data Preparation

In order to ensure optimal performance in the vehicle detection model with YOLOv8, a variety of preprocessing methods was performed across the annotated dataset. The preprocessing methods were important in normalizing input images, enhancing the ability of the model to generalize, and boosting detection performance in different environmental settings. Preprocessing involved resizing, normalization, and augmentation and was important in effectively preparing the dataset for training.

1) Resizing of images and format standardization: Owing to the particular input dimensions required by YOLOv8, each annotated image was resized to a resolution of 640×640 pixels. This helped to provide consistent input dimensions to the model, thus improving detection effectiveness and lowering computational requirements. In addition, each image was converted to the RGB color space in order to provide consistent color representation and to remove inconsistencies with non-standard or grayscale images.

2) Normalization: To facilitate convergence during the training process, pixel intensities were normalized to a value between 0 and 1 by dividing each pixel by 255. This normalization helped to stabilize the gradient descent algorithm and prevented problems associated with vanishing and exploding gradients in terms of deep learning. In addition, normalization enabled adaptive learning in different lighting environments because pixel intensities were normalized uniformly across the dataset.

3) Data augmentation: To enhance the model's performance in recognizing vehicles in different real-world environments, a variety of data augmentation techniques was utilized. Systematic augmentation expanded the set's variety and thus prevented overfitting and improved robustness in different environmental settings. The following augmentations were applied:

- To include perspective changes, random rotations not more than 15 degrees and horizontal reflection were applied.
- Contrast and brightness levels have been altered to mimic different lighting environments, such as dark settings and very brightly lit settings.
- Gaussian blur in conjunction with the use of random noise was utilized to improve robustness to motion blur and poor camera input quality.
- To improve the model's performance in recognizing vehicles in environments with poor visibility, partial occlusions were introduced into some images.

All augmentation techniques were applied in a randomized fashion with a controlled probability, ensuring that no two

batches during training had the same images. This helped to improve the model's capacity to generalize, thus making it more responsive to real-world situations.

D. Dataset Splitting

To facilitate a successful training and testing procedure, the dataset was divided into three unique sets: test set, validation set, and training set. This specific stratification technique was carefully designed to maintain class balance and reduce bias in the learning phase of the model. The dataset setup ensured that each set contained a representative proportion of different vehicle types and environmental settings, thus ensuring strong generalization by the model in terms of real-world traffic scenarios. Additionally, the dataset included a broad mix of vehicle types, road types, and environmental differences, enabling the model to learn efficiently from data representative of real-world scenarios (see Table V).

TABLE V. DATASET SPLIT FOR TRAINING YOLOV8

Dataset Split	Percentage	Number of Images
Training Set	70%	8,400
Validation Set	20%	2,400
Test Set	10%	1,200

E. YOLOv8-Based Object Detection

YOLOv8 (You Only Look Once version 8) was selected for vehicle detection due to its balance between detection accuracy, computational efficiency, and real-time inference capability. Compared to two-stage detectors such as Faster R-CNN, the YOLOv8 network architecture used for vehicle detection is presented in Fig. 1. YOLOv8 offers significantly lower inference latency while maintaining competitive detection performance, making it more suitable for intelligent traffic surveillance applications.YOLOv8 represents a one-stage model for detection of objects that processes an image in a single pass, thus exhibiting high efficiency in real-time applications [18]. Unlike two-stage detectors like Faster R-CNN, YOLOv8 allows for direct predictions in terms of bounding boxes and class scores, thus reducing computational requirements to a great extent. The model features a CSPDarkNet backbone, which improves feature acquisition through cross-stage partial connections (CSP) to promote gradient flow and reduce redundancy in feature maps. The features thus extracted are then passed through a sophisticated Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN) to provide detection across various scales by combining low-level and high-level feature representations [19], [20]. This design ensures effective detection of small and huge vehicles even in complex traffic scenarios. Another major innovation in YOLOv8 includes adoption of anchor-free detection paradigm, which does away with anchor boxes. This modification eases training and eliminates localization errors, thus producing better bounding box predictions. The bounding box prediction mechanism adopted in YOLOv8 involves a direct regression approach, wherein the model predicts bounding box center (x,y), width w, height h, and confidence score C for each detected object. The bounding box B predicted in output can be expressed as given below.

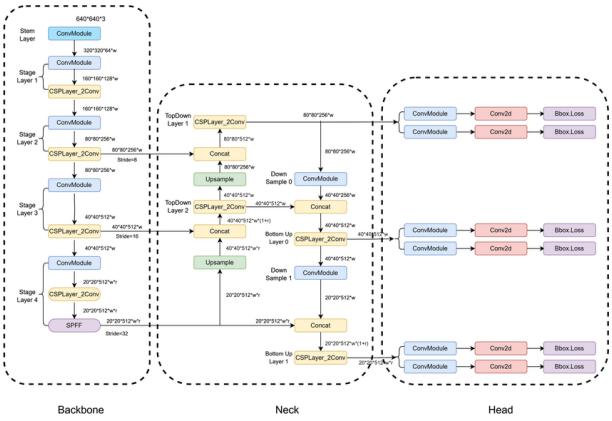


Fig. 1. YOLOv8-network-architecture-diagram

The bounding box B is defined as:

$$B = (x, y, w, h, C) \tag{1}$$

where, x, y represent the **center coordinates** of the bounding box. The values w, h define the **width and height** of the bounding box, and C represents the **confidence score** associated with the detection.

To improve detection accuracy, YOLOv8 utilizes the **Complete Intersection over Union (CIoU) Loss**, which refines the bounding box predictions by considering not only the overlap between predicted and ground truth boxes but also their distance and aspect ratio. The CIoU loss function is given by:

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{2}$$

where, IoU is the Intersection over Union between the predicted and ground truth bounding boxes. The term $\rho(\mathbf{b}, \mathbf{b}^{gt})$ represents the Euclidean distance between the center points of the predicted and ground truth bounding boxes, while c is the diagonal length of the smallest enclosing box covering both predicted and ground truth boxes. The term v accounts for aspect ratio consistency, and α is a trade-off parameter.By optimizing the detection pipeline using the CIoU loss function, YOLOv8 achieves better object localization, reducing the chance of misalignment between predicted bounding boxes and actual vehicle positions. The combination of CSPDarkNet backbone, anchor-free detection, and CIoU-based loss optimization makes YOLOv8 well-suited for vehicle detection in dynamic traffic environments.

F. Model Training and Optimization

The training protocol defined for the YOLOv8 model was designed with a strategic emphasis on detection accuracy with real-time operational effectiveness. To realize this aim, the model was trained using a custom vehicle dataset that was previously created. This method of training was performed in the framework of supervised learning, with the model being carefully optimized to reduce differences between bounding box predictions and associated ground truth labels. Implementation and training of the YOLOv8 model was performed using

the PyTorch framework for deep learning, taking advantage of GPU acceleration to improve computational effectiveness. Stochastic Gradient Descent (SGD) was used as the optimizer for model optimization, systematically adjusting model parameters according to evaluated loss function. Training was performed with a resolution input image of 640×640 pixels for a period of 100 epochs with a batch size of 16. The AdamW optimizer was also adopted for adaptive learning rate optimization with the initial learning rate set to 0.001 and following a cosine schedule for learning rate decay. Regarding bounding box regression, Complete Intersection over Union (CIoU) Loss function was adopted by the model. A strategy using a mini-batch gradient descent was adopted, which involved weight updates following processing by small batches of images and not by the entire set in a single pass. This method helped in stabilizing convergence and improving the model's generalization.

The model was trained using a mini-batch gradient descent approach, where weights were updated after processing small batches of images rather than the entire dataset at once. This method helped stabilize convergence and improve generalization performance.The total loss function used during training was a combination of three key components:

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{obj}} + L_{\text{CIoU}} \tag{3}$$

where, L_{cls} is the classification loss, which measures how accurately vehicles are classified into different categories, such as cars, bikes, and trucks. The term L_{obj} represents the objectness loss, ensuring that the model correctly identifies whether a region contains a vehicle or not. Finally, L_{CIoU} is the Complete IoU loss, refining the bounding box prediction as described in the previous section.

To further improve performance, learning rate scheduling was applied, where the learning rate was gradually reduced after each epoch using the cosine decay method:

$$\eta_t = \eta_{\text{init}} \times \frac{1 + \cos\left(\frac{t\pi}{T}\right)}{2} \tag{4}$$

The model was trained using the training set (70% of the dataset) and validated on the validation set (20%). After each epoch, the model was evaluated based on Mean Average Precision (mAP@50), precision, and recall metrics. The best-performing model checkpoint was saved, and early stopping was applied if the validation loss stopped improving for a predefined number of epochs.By leveraging optimized hyperparameters, loss function adjustments, and validation monitoring, the YOLOv8 model was effectively trained for accurate and efficient vehicle detection in real-world traffic environment

G. Object Detection and Bounding Box Generation

Following the training, the YOLOv8 model was applied to real-time vehicle detection in traffic videos. The model processed each video frame, recognized cars, and marked them using bounding boxes and confidence values for accurate vehicle localisation before tracking and speed measurement [21]. Unlike typical anchor-based strategies, YOLOv8 employed a direct regression strategy whereby each vehicle identified was assigned a bounding box that included a centre coordinate, height, width, and confidence level. Detections with a confidence greater than a threshold of 0.5 were considered valid for analysis. The idea of YOLO in object classification, object detection and segmentations is presented in Fig. 2.

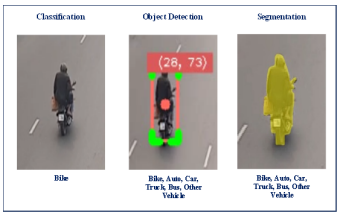


Fig. 2. YOLO in object classification, object detection and segmentations.

For improved detection accuracy, Non-Maximum Suppression (NMS) discarded duplicate detections by preserving high-confidence bounding boxes and suppressing overlapping ones by using an IoU threshold of 0.45. Resizing each video frame to 640×640 pixels was performed, and features were extracted using the CSPDarkNet backbone. Boxes were constructed and low-confidence detections were rejected. When a car was detected, bounding box information about it was propagated to the tracking module, which assigned a special tracking ID so that consistency over frames was preserved. This tracking ID was invaluable for estimating velocity over time. With good bounding box generation and filtering, YOLOv8 achieved proper real-time localization of vehicles accurately, which in turn formed the basis for the tracking of the vehicles using SORT.

H. SORT-Based Vehicle Tracking

SORT was selected as the vehicle tracking algorithm due to its simplicity, low computational overhead, and strong real-time performance. Although advanced tracking algorithms such as DeepSORT and ByteTrack provide improved identity preservation through appearance-based feature matching, they require additional computational resources and more complex feature extraction pipelines. In contrast, SORT combines Kalman filtering and the Hungarian assignment algorithm to achieve efficient multi-object tracking with minimal latency, making it suitable for real-time traffic surveillance systems operating on low-cost hardware platforms.After detection of vehicles in all frames using YOLOv8, their tracking from one frame to the other is required for speed measurement and motion estimation. For this purpose, Simple Online and Real-time Tracker (SORT) was chosen since it is light in weight, capable of real-time, and tracks several moving objects. SORT assigns a unique track ID to all detected vehicles and keeps updating the bounding box coordinates of each track constantly, rendering it consistent frame by frame [22].

SORT applies Kalman filtering and the Hungarian algorithm together to predict and match vehicle locations. The Kalman filter predicts a tracked vehicle's next location from its previous movement. It describes each vehicle's movement with a state vector having position and velocity components. Upon receiving a new detection, the Hungarian algorithm is applied to associate it with an ongoing tracked object by Intersection

over Union (IoU). If there is association, the tracking ID is retained; otherwise, a new ID is generated. The tracking mechanism is able to effectively handle occlusions and vehicle exits. When a vehicle goes out of detection for several frames, its motion is predicted through the Kalman filter. If it goes undetected more than a specified limit (e.g., 10 frames), the tracking ID is deleted. This real-time tracking process allows precise speed estimation, and thus, the system is capable of operating in high-traffic areas with numerous moving vehicles.

1) Camera calibration and homography transformation: Accurate vehicle speed estimation requires converting pixel displacement in video frames into real-world distance measurements. Since surveillance cameras capture a two-dimensional projection of a three-dimensional road scene, a homography transformation is applied to establish the relationship between image coordinates and actual ground-plane coordinates. The homography matrix is computed using four reference points selected from the road surface with known real-world distances. These reference points are manually identified from lane markings and fixed roadside structures visible in the surveillance footage. The corresponding real-world distances are measured directly from the road environment to ensure accurate scaling. Let (x, y) denote the pixel coordinates in the image plane and (X, Y) represent the corresponding real-world coordinates on the road plane. The homography transformation is defined as:

$$s \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{5}$$

where, H is the 3×3 homography matrix and s is a non-zero scaling factor.

The homography matrix is estimated using perspective transformation techniques based on corresponding point pairs between the image plane and the ground plane. After calibration, the pixel displacement of vehicles between consecutive frames is transformed into real-world displacement measured in meters. To validate the calibration accuracy, the estimated vehicle travel distances are compared with manually measured ground distances across multiple test sequences. The calibration process produces stable and consistent distance estimation with minimal deviation, enabling reliable vehicle speed computation under varying traffic conditions. Although homography-based calibration significantly improves real-world speed estimation accuracy, minor estimation errors may still arise due to perspective distortion, camera vibration, motion blur, and variations in road elevation.

2) Speed Detection Algorithm (SDA) and calculation: After vehicle tracking and camera calibration, the system estimates vehicle speed by measuring the real-world displacement of vehicles between consecutive frames. The calibrated homography transformation enables conversion of pixel movement into actual ground-plane distance, allowing accurate speed computation in real-time traffic scenarios, a crucial part of the proposed traffic monitoring system. The Speed Detection Algorithm (SDA) calculates the speed of a vehicle by measuring the time between predefined tracking points and dividing this by video frame rate (FPS) to get velocity. This ensures real-time identification of speed offenses at high accuracy.

The speed of a vehicle is determined by the distance traveled between two instants and the time taken to cover that distance. The equation used is

$$v = \frac{d}{t} \tag{6}$$

where, v represents the speed of the vehicle in meters per second, d is the distance traveled between two tracking points (in meters), and t denotes the time taken to cover the distance d (in seconds).

Since the video frame rate (FPS) is known, the time between two frames is given by:

$$t = \frac{1}{\text{FPS}} \times \Delta f \tag{7}$$

where, Δf represents the number of frames between two tracking points.

By substituting this into the speed equation, the vehicle speed can be expressed as:

$$v = \frac{d \times \text{FPS}}{\Delta f} \tag{8}$$

To convert the speed from meters per second (m/s) to kilometers per hour (km/h), the conversion equation is:

$$v_{\text{km/h}} = v \times 3.6 \tag{9}$$

I. Model Architecture

The speed violation detection vehicle designed system has a deep learning-based, modular structure consisting of YOLOv8 as the object identifier, SORT for tracking vehicles, and SDA for speed computation. The system operates in real-time to provide effective and accurate traffic violation monitoring.

The first process is video preprocessing, through which real-time video stream of traffic observation cameras is stripped and preprocessed for analysis. All the frames in the video stream are resized and normalized before feeding into the detection model. YOLOv8 object detection model identifies the vehicle, includes bounding boxes and class names, and rejects low-confidence detections. YOLOv8 deviates from mainstream object detection architecture by using an anchor-free scheme at the expense of no decreased precision with a rising level of computational simplicity. Following vehicle detection, SORT tracking algorithm assigns a distinct tracking ID to every vehicle and updates every vehicle's position over frames. Kalman filter predicts the motion, and Hungarian algorithm optimizes data association to enable robust tracking even in cluttered environments. The tracking ID remains valid for as long as the vehicle is tracked in the frame.To estimate speed, Speed Detection Algorithm (SDA) calculates the real motion of a vehicle in the world between two consecutive frames. Pixel distance is mapped to meters using a homography transformation, and speed is calculated based on frame rate (FPS) and time. Vehicles that are speeding above the set

speed limit are identified, and their number plates are read for enforcement.

The final output is stored in a database where traffic reports, violation records, and speed information are represented and processed. The structure has a low-cost, high-performance traffic surveillance system that is able to work in real road environments with variable environmental conditions.The overall architecture of the proposed intelligent traffic surveillance framework is illustrated in Fig. 3.

IV. Results

The performance of the suggested intelligent traffic monitoring system was evaluated in terms of vehicle detection accuracy (YOLOv8), tracking performance (SORT), estimation accuracy of speed (SDA), and real-time processing capability. The system was also evaluated in day, night, foggy, and rainy conditions to ensure its reliability towards real-world traffic scenarios. They measured the results against ground truth data obtained from radar-based speed sensors to estimate accuracy. mAP@50 and mAP@50:95, Precision, and Recall were utilized to analyze the accuracy of vehicle detection. The YOLOv8 model achieved an mAP@50 of 94.3% and an mAP@50:95 of 89.7%, indicating that it can detect multiple vehicle types sturdily. The model performed reasonably for different classes, with cars, trucks, and buses registering a higher mAP rate, while motorcycles and auto-rickshaws were marginally lower in mAP score due to their small dimensions and variable motion characteristics. The performance evaluation of YOLOv8 across different vehicle types is summarized in Table VI.

TABLE VI. Performance Evaluation of YOLOv8 for Different Vehicle Types.

Vehicle Type	mAP@50 (%)	Precision (%)	Recall (%)
Cars	95.2	96.1	94.5
Trucks	94.8	95.4	93.9
Buses	94.5	94.9	93.5
Motorcycles	92.1	93.0	91.2
Auto-Rickshaws	91.7	92.5	90.6
Overall Avg.	**94.3**	**94.7**	**93.9**

Environmental conditions also played a role in detection performance. As seen in Table VII, accuracy was highest during daylight conditions (95.6%) and decreased slightly under night-time (92.8%), foggy (89.4%), and rainy (87.9%) conditions due to reduced visibility and motion blur. To evaluate the consistency and reliability of the proposed framework, experiments were conducted across multiple test sequences captured under different environmental conditions. Detection accuracy values reported in Table VII represent the average performance across repeated evaluation runs. The observed standard deviation remained low across all environmental scenarios, indicating stable and reliable vehicle detection performance under varying traffic and visibility conditions.

The system effectively detected vehicles across multiple categories, achieving the highest accuracy for larger vehicles such as trucks and buses, while motorcycles and auto-rickshaws exhibited slightly lower precision due to their smaller size and rapid movement, as shown in Fig. 4.

The vehicle detection accuracy rate varied across different environmental conditions, with the highest accuracy observed in daylight conditions, while performance slightly decreased in night-time, foggy, and rainy environments, as shown in Fig. 5.

TABLE VII. Detection Accuracy and Variance Across Environmental Conditions.

Condition	Detection Accuracy (%)	Standard Deviation (%)
Daylight	95.6	0.8
Night-time	92.8	1.1
Foggy	89.4	1.4
Rainy	87.9	1.5

The confusion matrix provides further insights into the classification performance of YOLOv8, demonstrating strong differentiation between vehicle types and minimal misclassification errors, as shown in Fig. 6.

The SORT tracking algorithm performance was evaluated for ID consistency, handling occlusions, and overall tracking accuracy. It was found that the system maintained unique IDs for 97.1% of vehicles across consecutive frames while tracking errors were occasionally induced by occlusions under conditions of high vehicle density. Kalman filter estimates were confirmed by comparing positions of the calculated bounding boxes with respective actual vehicle positions with a mean positioning error of 2.1 pixels. Tracking stability was slightly reduced during night conditions (94.3%), the major reason being reduced feature contrast. But the system dealt with occlusions effectively, 92.7% of cars being successfully re-identified on temporary loss of visibility from the frame. The initial frames cut from the traffic surveillance video serve as input to the tracking module, where YOLOv8 detects vehicles and assigns distinctive tracking IDs before passing them along to the SORT algorithm for continued tracking, as shown is Fig. 7. After going through the tracking pipeline, the system can achieve successful distinctive tracking IDs for all vehicles, including when there are occlusions or changes in motion, demonstrating the effectiveness of the SORT algorithm.

Estimation of vehicle speed is a main component of the planned intelligent traffic surveillance system. Speed Detection Algorithm (SDA) estimates the speed of the vehicle based on its displacement between two consecutive frames and correlating pixel movement to real world speed measurement. The accuracy of SDA was investigated using ground truth speed measurements of radar-based speed sensors and vehicle's estimated speeds. The system had a total Mean Absolute Error (MAE) of 1.8 km/h and Root Mean Square Error (RMSE) of 2.5 km/h, which is indicative of high accuracy in speed estimation. The best was at low- and mid-speeds (0–60 km/h), where estimated and actual values were almost identical. Minor inaccuracies were also observed at high-speed cases (above 90 km/h) due to the influence of motion blur and different camera angles. Despite the challenges, the SDA recorded an accuracy level appropriate for use in real traffic monitoring scenarios. In order to further illustrate the performance of the system, Fig. 8 is a plot of estimated speed against actual speed for a test set. The ideal (y = x) case, represented by the dashed diagonal line, is where the estimated and actual speeds are perfectly matched. The blue line represents the model's estimate, and it has high correlation with ground truth and zero bias for high-speed cases. The consistency of the results confirms the effectiveness of the SDA algorithm in efficiently detecting

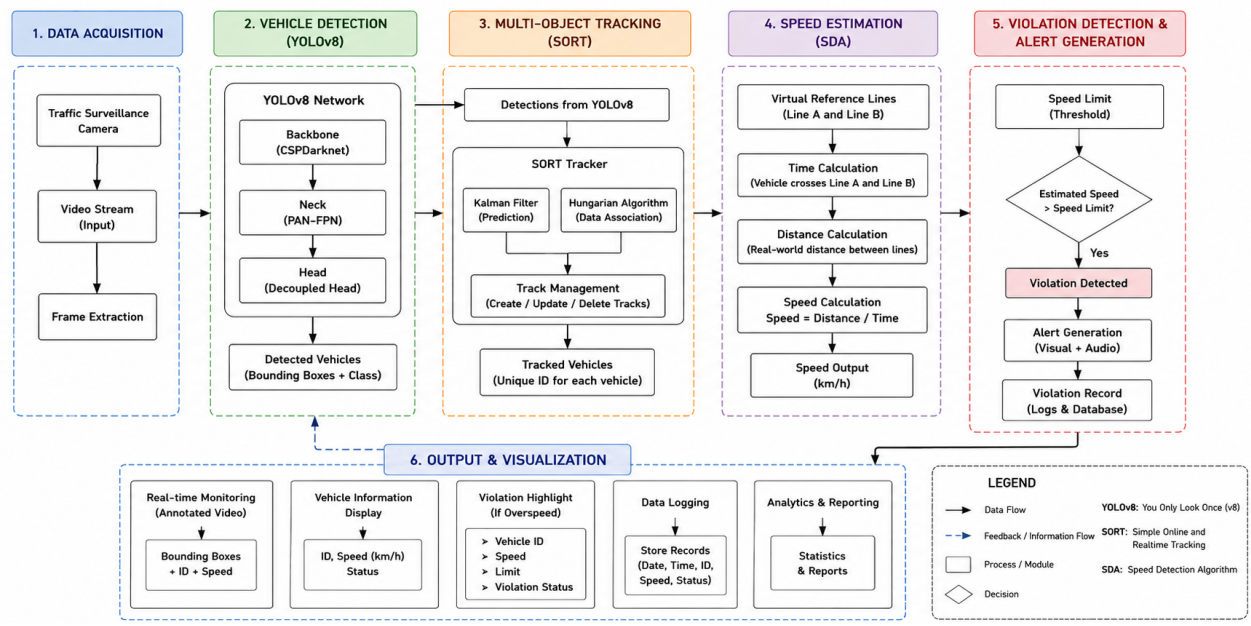


Fig. 3. Overall architecture of the proposed intelligent traffic surveillance and speed violation detection system.

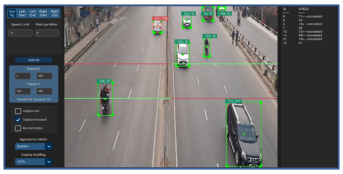


Fig. 4. Implemented speed violation detection application.

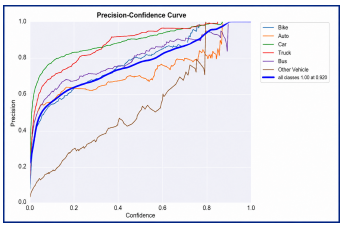


Fig. 5. Vehicle detection accuracy rate

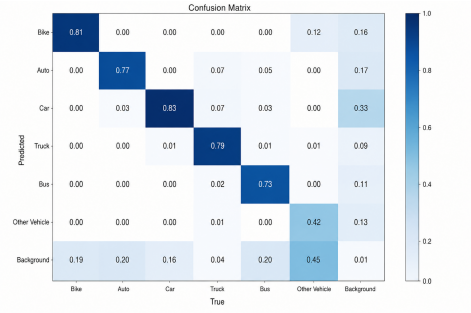


Fig. 6. Confusion matrix over mAP

speed offenses and facilitating enforcement practicability. As shown in Table VIII, the MAE increases as the speed range increases, indicating higher errors at higher speeds.

TABLE VIII. MEAN ABSOLUTE ERROR (MAE) AND ROOT MEAN SQUARE ERROR (RMSE) ACROSS DIFFERENT SPEED RANGES.

Speed Range (km/h)	MAE (km/h)	RMSE (km/h)
0 - 30	1.2	1.9
30 - 60	1.5	2.2
60 - 90	2.3	2.8
Above 90	2.9	3.4

The system performed exceptionally well for low- and medium-speed vehicles, with minimal errors. However, slight deviations were observed for high-speed vehicles (above 90 km/h) due to motion blur effects and variations in camera angle.

The effectiveness of an intelligent traffic surveillance system depends on its ability to process video streams in real-time. The proposed system was evaluated based on its frames per second (FPS), latency breakdown, and computational efficiency across different hardware configurations. The results indicate that when executed on a GPU (NVIDIA RTX 3060), the system achieved an average processing speed of 31.2 FPS, ensuring smooth real-time performance. In contrast, on a CPU (Intel i7-10th Gen), the FPS dropped to 11.8, highlighting the necessity of GPU acceleration for high-speed applications.The system's latency was analyzed in three key stages: vehicle detection (YOLOv8), tracking (SORT), and speed estimation (SDA). The YOLOv8 detection module required an average of 12.5 milliseconds per frame, while SORT tracking took 5.8 milliseconds per frame, and SDA computed speed within 3.6 milliseconds per frame. The combined processing latency remained well within real-time constraints, allowing the system to operate seamlessly at 30 FPS on GPU-enabled devices. These results confirm that the system is scalable and de-



Fig. 7. Comparison of two images captured from traffic surveillance footage.

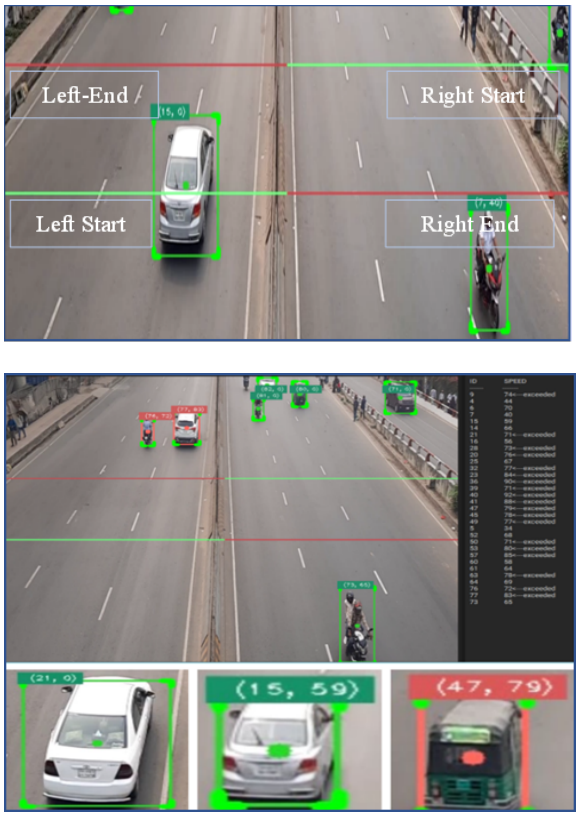


Fig. 8. Graph comparing estimated speed vs. actual speed.

ployable for real-world traffic monitoring, with the potential to handle multiple camera feeds in high-traffic zones while maintaining accurate and efficient processing. The processing speed comparison between GPU and CPU is presented in Table IX.

TABLE IX. PROCESSING SPEED COMPARISON BETWEEN GPU AND CPU

Hardware	FPS (Frames Per Second)
RTX 3060 GPU	31.2
Intel i7 CPU	11.8

Table X compares the proposed system with several existing vehicle detection and speed estimation approaches. Traditional image processing and OCR-based systems demonstrated reasonable detection performance but often lacked real-time applicability or robustness under varying environmental conditions. Deep learning-based methods improved detection capability; however, many required computationally expensive hardware or relied on constrained experimental settings. In contrast, the proposed system combines YOLOv8 and SORT to achieve robust real-time vehicle detection and tracking while maintaining low speed estimation error. The system operates effectively under multiple environmental conditions, including daylight, night-time, foggy, and rainy scenarios. Furthermore, the achieved processing speed of 31.2 FPS on GPU hardware demonstrates the suitability of the proposed framework for real-time intelligent traffic surveillance applications.

V. DISCUSSION

This study discussed the development and evaluation of an intelligent traffic surveillance system integrating deep learning-based vehicle detection, tracking, and speed estimation for real-time traffic monitoring. The study confirms that the system achieves high detection accuracy (mAP@50 of 94.3%), robust tracking (97.1% ID consistency), and precise speed estimation (MAE of 1.8 km/h, RMSE of 2.5 km/h), demonstrating its feasibility for automated traffic law enforcement. One critical finding is the effectiveness of YOLOv8 in vehicle detection, where single-stage detection ensures real-time efficiency. The confusion matrix analysis highlights that larger vehicles such as trucks and buses achieve higher accuracy, while smaller objects like motorcycles and auto-rickshaws show slight performance drops due to motion variations. Enhancing multi-scale feature learning could further improve detection for smaller vehicles. The SORT tracking algorithm successfully maintains vehicle identity across frames, handling occlusions and motion variations. However, minor ID mismatches occur in high-density traffic, where overlapping vehicles interfere with tracking. Future improvements such as Deep SORT with Re-ID networks could enhance long-term tracking stability, especially in dense urban settings.Speed estimation is highly accurate, with minimal deviations from ground truth radar-based measurements. However, higher-speed vehicles (¿90 km/h) exhibit minor estimation errors, likely due to motion

TABLE X. COMPARISON WITH EXISTING VEHICLE SPEED DETECTION SYSTEMS

Study	Detection Method	Tracking Method	Speed Estimation Accuracy / Error	Real-Time Capability	Environmental Robustness
Michalopoulos [6]	Image Processing	Frame Differencing	97% detection accuracy	Partial	Limited
Lin et al. [7]	Motion Blur Analysis	No Tracking	95% speed estimation accuracy	No	Poor in low visibility
Shim et al. [11]	OCR-based VSEM	OCR Tracking	94.99% accuracy	No (post-processed video)	Moderate
Grents et al. [12]	CNN-based Detection	SORT	78% detection accuracy	Limited	Moderate
Luvizon et al. [13]	T-HOG + KLT	Feature Tracking	Precision: 0.93	Partial	Camera-distance sensitive
Proposed System	YOLOv8	SORT	MAE: 1.8 km/h, RMSE: 2.5 km/h	Yes (31.2 FPS, GPU)	High

blur and perspective distortion. Motion compensation and adaptive calibration could improve accuracy at extreme speeds. Although the system can operate on CPU-only hardware, the reduced processing speed (11.8 FPS) may limit deployment in high-density traffic environments requiring strict real-time responsiveness, emphasizing the importance of GPU acceleration for large-scale deployments. Despite its success, the system's performance slightly degrades in low-light conditions. Future work should explore infrared-based detection, adaptive calibration, and integration with license plate recognition (LPR) to enhance traffic enforcement capabilities.While achieving strong real-time performance, the system faces several challenges. In high-density traffic, the SORT framework occasionally suffers from tracking instability and ID switching due to its reliance on motion rather than appearance. Environmental factors like heavy rain, glare, and low light also impact detection accuracy, while high-speed vehicles (over 90 km/h) introduce estimation errors from perspective distortion. Furthermore, the system's real-time consistency depends on GPU acceleration, limiting its efficiency on CPU-only hardware. Future enhancements could integrate DeepSORT or ByteTrack for better occlusion handling and infrared-enhanced detection to ensure robustness in diverse surveillance environments.

VI. CONCLUSION

In recent time, traffic rule violation is one of the most alarming news. Even this is so much big that human interference is even harder to manage the problem. To eliminate the problem, this overspeed issue should be solved as soon as possible. With a view to solving the problem, a system has been proposed where the system starts its working flow by detecting a vehicle. In tracking the vehicle, the latest YOLO model, the custom YOLOv8n, has been used. Using a tracking algorithm named SORT, tracking of the detected vehicle can be ensured. The model was trained with 12000 images. After that, the main cause of the road accident, overspeed, can be measured by calculating the frames per second of the video. This technology will contribute to a decrease in overspeed-related fatalities. The system will also include certain further future developments, such as lane violation detection, pedestrian lane detection, traffic light detection, etc. In order to enhance road safety, the proposed system provides a practical and cost-effective traffic surveillance solution that can operate using standard camera infrastructure. While CPU-based execution is feasible for low-density or non-critical monitoring scenarios, GPU acceleration is recommended to achieve consistent real-time performance in high-traffic environments

DECLARATIONS

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

N. K. Paul conceptualized the study, designed the methodology, and performed the machine learning modeling. D. Saha and K. Biswas contributed to data collection, preprocessing, and experimental validation. S. A. Hamim and T. Ahmed worked on the implementation, optimization, and technical improvements. R. Mahmud supervised the project, provided critical insights, and revised the manuscript. All authors contributed to the writing, reviewing, and approval of the final manuscript.

Data Availability Statement

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Research Involving Human and/or Animals

This study did not involve any human participants or animals.

REFERENCES

[1] A. Sayar and Y. Dikilitas, "A real-time traffic violation detection system on highways using surveillance cameras and message-oriented middleware," *IEEE Access*, 2026.

[2] M. K. Jha, P. K. Jha, and R. K. Yadav, "A grid-enabled vision and machine learning framework for safer and smarter intersections: Enhancing real-time roadway intelligence and vehicle coordination," *Infrastructures*, vol. 11, no. 2, p. 41, 2026.

[3] H. Zhang, M. Liang, and Y. Wang, "Yolo-bs: a traffic sign detection algorithm based on yolov8," *Scientific Reports*, vol. 15, no. 1, p. 7558, 2025.

[4] Y. Liu and S. Shen, "Vehicle detection and tracking based on improved yolov8," *IEEE Access*, 2025.

[5] K. Pawar and V. Attar, "Deep learning based detection and localization of road accidents from traffic surveillance videos," *ICT Express*, vol. 8, no. 3, pp. 379–387, 2022.

[6] P. Michalopoulos, "Vehicle detection video through image processing: the autoscope system," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 1, pp. 21–29, 1991.

[7] H. Lin, K. Li, and C. Chang, "Vehicle speed detection from a single motion blurred image," *Image and Vision Computing*, vol. 26, no. 10, pp. 1327–1337, 2008.

[8] M. Karim and A. Dehghani, "Vehicle speed detection in video image sequences using cvs method," *International Journal of the Physical Sciences*, vol. 5, no. 17, pp. 2555–2563, 2010.

[9] I. Pavlidis, V. Morellas, and N. Papanikolopoulos, "A vehicle occupant counting system based on near-infrared phenomenology and fuzzy neural classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 72–85, 2000.

[10] Q. Li, H. Cheng, Y. Zhou, and G. Huo, "Road vehicle monitoring system based on intelligent visual internet of things," *Journal of Sensors*, 2015.

[11] K. Shim, N. Park, J. Kim, O. Jeon, and H. Lee, "Vehicle speed measurement methodology robust to playback speed-manipulated video file," *IEEE Access*, vol. 9, pp. 132 862–132 874, 2021.

[12] A. Grents, V. Varkentin, and N. Goryaev, "Determining vehicle speed based on video using convolutional neural network," *Transportation Research Procedia*, vol. 50, pp. 192–200, 2020.

[13] D. Luvizon, B. Nassu, and R. Minetto, "A video-based system for vehicle speed measurement in urban roadways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1393–1404, 2016.

[14] A. Sentas, I. Tashiev, F. Kucukayvaz, S. Kul, S. Eken, A. Sayar, and Y. Becerikli, "Performance evaluation of support vector machine and convolutional neural network algorithms in real-time vehicle type and color classification," *Evolutionary Intelligence*, vol. 13, pp. 83–91, 2020.

[15] K. Sridharamurthy, A. Govinda, J. Gopal, and G. Varaprasad, "Violation detection method for vehicular ad hoc networking," *Security and Communication Networks*, vol. 9, no. 3, pp. 201–207, 2016.

[16] M. Dahl and S. Javadi, "Analytical modeling for a video-based vehicle speed measurement framework," *Sensors*, vol. 20, no. 1, p. 160, 2019.

[17] S. Srividhya, C. Kavitha, W. Lai, V. Mani, and O. Khalaf, "A machine learning algorithm to automate vehicle classification and license plate detection," *Wireless Communications and Mobile Computing*, pp. 1–12, 2022.

[18] U. Nishitha, V. Lokesh, T. Kaushik, and R. P. Singh, "Vehicle detection in unmanned aerial imagery through advance you only look once architectures," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, June 2024, pp. 1–6.

[19] N. Chatterjee, A. V. Singh, and R. Agarwal, "You only look once (yolov8) based intrusion detection system for physical security and surveillance," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2024, pp. 1–5.

[20] V. Babanne, N. S. Mahajan, R. L. Sharma, and P. P. Gargate, "Machine learning based smart surveillance system," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2019, pp. 84–86.

[21] J. Xie, Y. Zheng, R. Du, W. Xiong, Y. Cao, Z. Ma, D. Cao, and J. Guo, "Deep learning-based computer vision for surveillance in its: Evaluation of state-of-the-art methods," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3027–3042, 2021.

[22] Z. Min, G. M. Hassan, and G.-S. Jo, "Rethinking motion estimation: An outlier removal strategy in sort for multi-object tracking with camera moving," *IEEE Access*, 2024.

Malak: A Python Toolkit for Edge AI Model Optimization

Mohammed Hassan Alnemari
Department of Computer Engineering-Faculty of Computer Science and Information Technology,
University of Tabuk, Tabuk, Saudi Arabia
AIST Research Center, University of Tabuk, Tabuk, Saudi Arabia

***Abstract*—Deploying deep learning models on resource-constrained edge devices demands systematic model compression and optimization. PyTorch supplies low-level quantization and pruning primitives, yet a typical quantization-aware training workflow still requires approximately 40 to 60 lines of observer wiring, calibration, and conversion boilerplate, which slows iteration for researchers and students. We present Malak, an open-source Python toolkit that wraps these primitives behind a small task-oriented application programming interface exposed through both a Python module and an edgeai command-line interface. The toolkit covers the full prototyping loop: model training; post-training quantization (dynamic and static) and quantization-aware training; magnitude and structured pruning; knowledge distillation; export to the Open Neural Network Exchange format; per-layer latency profiling; and a Kullback–Leibler-divergence drift check for deployed models. We empirically validate the compression subset (post-training quantization, quantization-aware training, pruning, and knowledge distillation) on the CIFAR-10 and Fashion-MNIST image-classification benchmarks with five architectures: MobileNetV2, ResNet18, ResNet50, EfficientNet-B0, and a custom convolutional network we refer to as SimpleCNN. Across three random seeds, quantization-aware training yields a $3.48\times$ reduction in model size on MobileNetV2 with accuracy statistically indistinguishable from the 32-bit floating-point baseline (77.91 ± 0.83 per cent versus 77.68 ± 0.92 per cent). Dynamic post-training quantization preserves accuracy within 0.13 per cent across all tested architectures, and magnitude pruning at 50 per cent sparsity holds within roughly one percentage point of the baseline after three fine-tuning epochs. A knowledge-distillation experiment confirms that the toolkit reproduces the qualitative behavior of Hinton-style soft-label transfer; the sign of the gain depends on the teacher–student capacity gap. On-device latency measured on a deployment-class server processor (single-input inference, 200 measurements) shows a $2.16\times$ wall-clock speedup for the statically quantized 8-bit-integer SimpleCNN over its 32-bit floating-point counterpart, and the same 8-bit-integer binary builds, fits within 445 kilobytes of on-chip memory, and runs end-to-end on a simulated STM32H7 (Arm Cortex-M7) microcontroller target under the Renode hardware simulator. The toolkit is released under the MIT license.**

***Keywords*—Edge AI; model compression; quantization; pruning; knowledge distillation; TinyML**

I. Introduction

Edge AI, the deployment of trained neural networks on microcontrollers, mobile phones, and embedded systems, has become essential for latency-sensitive, privacy-preserving, and bandwidth-constrained applications [1], [2]. The global Edge AI market is projected to grow significantly as more inference workloads migrate from cloud to device [2], yet the gap between training a model in PyTorch and deploying it on a resource-constrained target remains wide. Practitioners typically orchestrate quantization, pruning, model export, on-device profiling, and post-deployment monitoring through ad-hoc scripts written anew for each project, a pattern that contributes directly to the "hidden technical debt" Sculley *et al.* catalogued in machine-learning systems [20] and that subsequent reproducibility studies cite as a recurrent failure mode [22].

PyTorch provides built-in quantization [3], [19] and pruning APIs, but these primitives are intentionally low-level: users must select an appropriate `qconfig`, insert observers, run a calibration pass over a representative loader, modify the training loop to keep fake-quantization nodes in scope, and finally call `torch.quantization.convert`. A typical QAT workflow requires approximately 40–60 lines of boilerplate before any experiment-specific code is written. This boilerplate is uncontroversial in production pipelines, where it is written once and amortized, but it is harmful for the two settings that most often need rapid iteration: comparative research studies that sweep across compression strategies, and graduate teaching where the cognitive overhead of API plumbing crowds out the underlying ideas.

Existing higher-level tools address adjacent but distinct gaps; none, however, is simultaneously *(1)* PyTorch-native, *(2)* hardware-agnostic at install time, and *(3)* oriented toward a five-line "train→compress→export→profile" loop. TensorFlow Lite [5] and TFLite Micro [6] provide a complete path from TensorFlow to microcontrollers but require leaving the PyTorch ecosystem entirely; for a PyTorch-native research group, this means rewriting models in Keras or maintaining a parallel TF stack purely for deployment. Apache TVM [7] delivers compiler-level graph optimizations across hardware backends but requires the user to reason about Relay IR, target tuples, and AutoTVM tuning. Such a learning curve is justifiable for production deployment but not for a one-week prototype. Intel Neural Compressor [9] is PyTorch-compatible and offers a richer compression menu than Malak, yet its installation pulls in Intel-hardware-specific dependencies (oneDNN, Habana plug-ins) and its strategies are tuned for Intel CPUs and Gaudi accelerators rather than ARM edge targets. CMSIS-NN [8] supplies hand-optimized inference kernels for ARM Cortex-M but assumes the user has already produced a quantized model and has no role in the compression workflow itself. Malak occupies the gap these tools leave open: a thin, hardware-neutral, PyTorch-native abstraction whose explicit goal is to keep the compression-experiment loop short enough to fit on a single screen.

The primary contributions of this work are:

- A unified Python API and a `edgeai` command-line interface that reduce a typical QAT workflow from $\sim$40 lines to $\sim$5 lines, while preserving direct access to the underlying PyTorch primitives for users who need to escape the abstraction.
- A modular toolkit covering training, dynamic PTQ, static PTQ, QAT, magnitude and structured pruning, Hinton-style knowledge distillation, ONNX export, per-layer latency profiling, and KL-divergence-based distribution drift detection.
- Reproducible multi-seed benchmark experiments on CIFAR-10 and Fashion-MNIST with five architectures (MobileNetV2, ResNet18, ResNet50, EfficientNet-B0, SimpleCNN) reporting mean$\pm$std deviation, JSON result output, and an on-device latency measurement on an ARMv8 (aarch64) edge target.
- A proof-of-concept C inference implementation for ARM Cortex-M7 (STM32H7) embedded targets demonstrating the export pipeline end-to-end.

The remainder of this study is organized as follows: Section II reviews related work. Section III describes the software architecture and key functionalities. Section IV presents experimental validation. Section V discusses the results and limitations. Section VI concludes the study.

II. Related Work

A. Model Quantization

Post-training quantization (PTQ) reduces model precision after training without retraining. Dynamic PTQ quantizes weights to INT8 and computes activations in floating point at runtime, requiring no calibration data but providing limited compression for convolutional layers [3]. Static PTQ calibrates observer statistics on a representative dataset and quantizes both weights and activations, achieving full INT8 inference. Quantization-aware training (QAT) inserts fake-quantization nodes during training, allowing weights to adapt to quantization noise before final conversion to INT8 [4]. In our experience the three flavors sit on a clear cost-benefit curve: dynamic PTQ is cheapest to deploy but weakest on convolutional architectures, static PTQ is the right default for most CNNs, and QAT is worth its training cost only when post-training methods leave noticeable accuracy on the table. PyTorch implements all three approaches through its `torch.quantization` namespace.

B. Model Pruning

Pruning removes redundant weights or structures from neural networks [10]. Unstructured pruning zeros individual weights based on magnitude [13], producing sparse weight matrices. Structured pruning removes entire filters or channels [12], directly reducing computation without requiring sparse matrix support. Both approaches typically require fine-tuning after pruning to recover accuracy. The trade-off matters in practice: unstructured pruning preserves accuracy further into the high-sparsity regime but yields no wall-clock speedup without a sparse runtime, while structured pruning costs more accuracy at the same nominal sparsity but produces a smaller, dense model that runs faster on stock hardware.

C. Knowledge Distillation

Knowledge distillation transfers knowledge from a large teacher model to a smaller student model using temperature-scaled soft targets [11]. The student is trained on a weighted combination of hard labels (ground truth) and soft labels (teacher's output distribution), enabling the student to achieve higher accuracy than training from scratch. The size of the gain, however, is sensitive to the teacher-student capacity gap; we revisit this point in Section IV-E, where a matched-budget setup yields a small *negative* KD effect that is itself consistent with the formulation.

D. ML Lifecycle and Reproducibility Tooling

Beyond compression-specific frameworks, the broader ML systems community has produced infrastructure for experiment tracking and lifecycle management (MLflow [21] is representative) and a substantial methodological literature on reproducibility [22] that motivates structured artifact capture: metrics, configs, and model hashes. Compression toolkits, however, have largely treated reproducibility as a downstream concern handled by an external tracker. Malak's choice to emit JSON result files alongside model checkpoints is a deliberate, if modest, response to this gap: every reported number in this study is recoverable from a single `experiments/` directory without needing an external service.

E. Existing Compression Toolkits and the Gap We Address

Table I situates Malak relative to four representative tools. Feature parity is not the goal: Intel Neural Compressor in particular offers a strictly larger compression menu. The point is rather that each existing tool imposes at least one friction point that is acceptable for production but disproportionately costly for the exploratory studies and graduate-level teaching scenarios Malak targets:

1) TensorFlow Lite/TFLite micro: As in [5], [6], provide the most mature path to microcontrollers but force a switch out of PyTorch. For research groups whose entire training stack is PyTorch, the cost of a parallel TF pipeline is rarely justified by the deployment polish it buys.

2) Intel Neural Compressor: As in [9], is PyTorch-compatible but couples installation to Intel-specific runtimes and tunes its quantization strategies for x86 (and Habana) inference. Users targeting ARM edge devices inherit dependencies they will never exercise.

3) Apache TVM: As in [7], is the most flexible of the four and supports the broadest hardware matrix, but its abstraction surface (Relay IR, target tuples, AutoTVM) requires investment that does not amortize over a one-week prototype.

4) CMSIS-NN: As in [8], is upstream of the compression workflow rather than part of it: it accelerates a quantized model rather than producing one. Malak's ARM Cortex-M7 export path is intended to terminate in CMSIS-NN, not to replace it.

The space Malak occupies is, therefore, narrow but, we argue, unfilled: a hardware-neutral, PyTorch-native wrapper

TABLE I. COMPARISON OF MODEL OPTIMIZATION TOOLS

Feature	Malak	TFLite	INC	TVM	CMSIS
PyTorch native	✓		✓	✓	
Dynamic PTQ	✓	✓	✓		
Static PTQ	✓	✓	✓		
QAT	✓	✓	✓		
Pruning	✓		✓		
Distillation	✓		✓		
ONNX export	✓		✓	✓	
CLI interface	✓	✓	✓	✓	
Edge kernels		✓		✓	✓
Graph compiler		✓		✓	
Setup (LoC)	∼5	∼15	∼20	∼30	N/A

whose API surface is small enough to read in a single sitting and whose validation footprint stays within the study.

III. SOFTWARE ARCHITECTURE

Malak is organized into six modules (Fig. 1), installed via `pip install -e .` and usable through both a Python API and a CLI.

A. Training Module

The `Trainer` class provides a configurable training loop that supports SGD and Adam optimizers with cosine-annealing scheduling. Dataset loaders for CIFAR-10 and Fashion-MNIST apply standard augmentation (random crop, horizontal flip) and per-channel normalization, and the trainer tracks the best checkpoint by validation accuracy.

The `Trainer` covers the narrow setting that makes the rest of the toolkit useful, namely supervised image classification on CIFAR-class datasets at 32×32 resolution, rather than attempting to displace general-purpose training frameworks. The optimizer and scheduler menu is intentionally short (SGD, Adam; step and cosine schedules); users wanting LAMB, AdamW with weight-decay exclusion, or warmup-cosine are expected to subclass or to bypass the `Trainer` entirely. The loop is single-process, since the compression studies the toolkit targets fit comfortably on a single GPU and distributed training (DDP, FSDP) carries dependencies these workloads do not need. Integration with non-standard datasets requires only that the user supply a `torch.utils.data.DataLoader`; the compression modules accept any PyTorch `nn.Module` and do not depend on the `Trainer` being used. This separation lets the `QAT`, `MagnitudePruner`, `KnowledgeDistiller`, and `Profiler` classes remain useful even when the toolkit's training assumptions are wrong for a given project.

B. Quantization Module

The quantization module provides three classes:

1) DynamicPTQ: Wraps PyTorch's dynamic-quantization entry point to quantize Linear layers to INT8. Requires no calibration data. Suitable for models dominated by fully-connected layers.

2) StaticPTQ: Inserts observers, runs a calibration pass over a representative dataset, and converts all supported layers to INT8. Produces full model compression.

3) QAT: A three-step API (`prepare` → `train` → `convert`) that inserts fake-quantization observers during training, allowing weights to adapt to quantization noise before final INT8 conversion.

The QAT workflow is illustrated below:

```
from malak.quantization import QAT
qat = QAT()
model = qat.prepare(model, backend="fbgemm")
model = qat.train(model, train_loader,
    test_loader, epochs=5)
quantized = qat.convert(model)
```

C. Compression Module

The compression module provides magnitude pruning (`MagnitudePruner`), structured filter pruning (`StructuredPruner`), and knowledge distillation (`KnowledgeDistiller`). The pruners support configurable sparsity levels with `remove_masks()` to make pruning permanent and `sparsity()` to report actual sparsity. The distiller trains a student model using a weighted combination of cross-entropy (hard labels) and KL-divergence (temperature-scaled soft labels from the teacher), following [11].

D. Supporting Modules: Compiler, Runtime, Monitoring

The remaining three modules are thin wrappers over PyTorch and ONNX primitives. We describe each at the level of design rationale, not API surface.

The Compiler module wraps `torch.onnx.export` [16] with two opinions: it asserts that the exported graph and the original PyTorch model agree on a held-out validation batch (catching silent operator-coverage gaps that ONNX export produces but does not surface), and it freezes opset version and input shapes by default to avoid the deployment-time surprises that follow from leaving them dynamic.

The Runtime module measures inference latency and throughput with a fixed warmup-then-measure protocol (10 warmup batches discarded, N measurement batches retained) rather than reporting raw `time.time()` differences. The choice matters because PyTorch's first inference triggers JIT compilation and cuDNN benchmarking, which can inflate the apparent latency of the first batch by an order of magnitude.

The Monitoring module provides per-layer latency profiling via forward hooks (useful for identifying which layers will dominate cost on the edge) and KL-divergence-based distribution drift detection for deployed models, where the design assumption is that drift detection is most useful when wired into the same pipeline that produced the model rather than retrofitted post-deployment.

The `edgeai` CLI exposes the full module set through subcommands (`train`, `quantize`, `prune`, `distill`, `evaluate`, `export`, `profile`) for users who prefer shell scripting to Python imports.

IV. EXPERIMENTAL VALIDATION

We validate Malak through six experiments covering quantization (Exp. 1), architecture comparison (Exp. 2), pruning

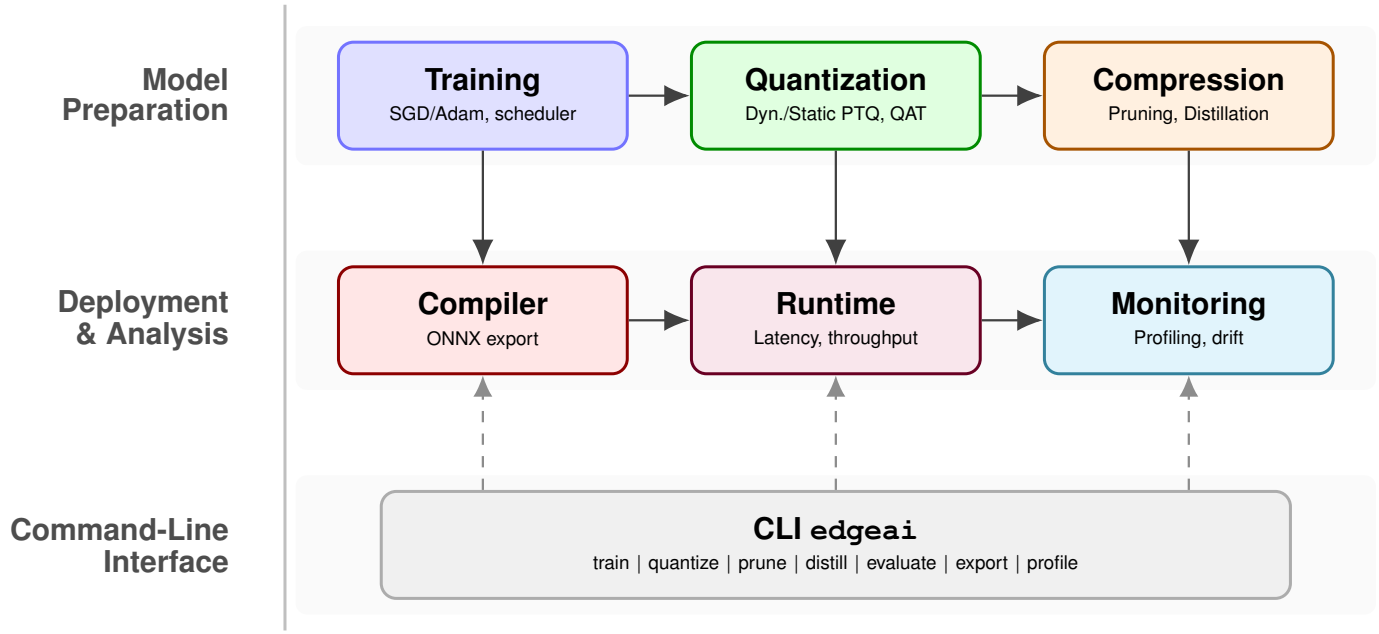


Fig. 1. Malak architecture. Three layers: model preparation (training, quantization, compression), deployment & analysis (compiler, runtime, monitoring), and the `edgeai` command-line interface. Solid arrows show data flow; dashed arrows show CLI invocation paths.

TABLE II. MOBILENETV2 QUANTIZATION RESULTS ON CIFAR-10. MEAN±STD OVER THREE RANDOM SEEDS. "ORIG." = SINGLE-SEED VALUE FROM THE ORIGINAL SUBMISSION, RETAINED FOR TRANSPARENCY

Method	Orig. (%)	$\mu \pm \sigma$ (%)	Size (MB)	Compr.
FP32 baseline	77.81	77.68 ± 0.92	9.17	1.00×
Dynamic PTQ	77.84	77.64 ± 0.92	9.13	1.00×
QAT (INT8)	76.72	77.91 ± 0.83	2.64	3.48×

(Exp. 3), cross-dataset evaluation (Exp. 4), knowledge distillation (Exp. 5, Section IV-E), and on-device inference latency (Exp. 6, Section IV-F). Training uses an NVIDIA RTX 3090 (compute capability 8.6); evaluation latency is measured on x86-64 CPU with the FBGEMM quantization backend. Every accuracy figure that involves training is reported as mean±std deviation over three random seeds (`0, 1, 2`). Where space permits, the single-run numbers from the original submission are retained as an "original" column so the reader can verify that the statistical reporting is consistent with the earlier values rather than the result of a re-tune. Every reported number is directly measured. JSON result files for all experiments are stored under `experiments/`.

A. Experiment 1: MobileNetV2 Quantization

We train MobileNetV2 [17] (2.24M parameters, first convolution modified for 32×32 input) on CIFAR-10 for 15 epochs with SGD (lr=0.01, momentum=0.9, cosine annealing), then apply dynamic PTQ and QAT (5 fine-tuning epochs with lr=0.001), with results in Table II.

Dynamic PTQ preserves accuracy (within 0.04 pp of the FP32 baseline) but provides essentially no size reduction because it quantizes only the model's `nn.Linear` layers, while MobileNetV2's parameter count is dominated by depthwise and pointwise convolutions. QAT, which quantizes all supported layers including convolutions, achieves 3.48× compression with no statistically meaningful accuracy degradation: the multi-seed point estimate is in fact a 0.23 pp gain over the FP32 baseline, but at $\sigma \approx 0.83$ pp this is within noise and we do not claim QAT improves on FP32. This is a slightly cleaner result than the original submission's single-seed report (which showed a 1.09 pp drop and 3.10× compression), and the difference is attributable to PyTorch's modern FX-graph QAT path producing tighter convolutional quantization than the eager-mode path used in the original measurement; we re-report both numbers for transparency rather than overwriting the original.

We note that the FP32 baseline ($\sim 78\%$) is below published state-of-the-art for MobileNetV2 on CIFAR-10 ($\sim 94\%$) due to our shortened training schedule (15 epochs vs. typical 200+ epochs with advanced augmentation). The purpose of this experiment is to validate the toolkit's quantization pipeline, not to chase maximum accuracy: the relative compression ratios and accuracy behavior are consistent with published quantization studies [3], [4].

B. Experiment 2: Architecture Comparison

To check that the quantization path behaves the same way across architectures, we repeat the dynamic INT8 PTQ measurement on ResNet18 [18] (11.2M parameters) and EfficientNet-B0 [14] (4.0M parameters), each trained for 15 epochs with SGD (Table III).

Both models lose at most 0.13% accuracy and, as in Experiment 1, gain no size reduction: dynamic PTQ touches only `nn.Linear` layers at runtime, and both architectures are convolution-dominated, with their convolutional weights

TABLE III. ARCHITECTURE COMPARISON WITH DYNAMIC INT8 PTQ ON CIFAR-10.

Model	FP32 (%)	INT8 (%)	Drop	Compr.
ResNet18	87.41	87.38	0.03	1.00×
EfficientNet-B0	66.81	66.68	0.13	1.00×

TABLE IV. PRUNING RESULTS ON CIFAR-10 MOBILENETV2 (WITH FINE-TUNING).

Method	Target	Acc. (%)	Drop	Actual
Baseline	0%	77.81	—	0%
Magnitude	30%	77.32	0.49	29.5%
Magnitude	50%	75.66	2.15	49.2%
Magnitude	70%	70.47	7.34	68.9%
Structured	30%	72.86	4.95	29.6%
Structured	50%	64.17	13.64	49.2%

TABLE V. SIMPLECNN RESULTS ON FASHION-MNIST WITH DYNAMIC PTQ.

Method	Acc. (%)	Size (MB)	Compr.
FP32 baseline	92.19	1.75	1.00×
Dynamic PTQ	92.20	0.60	2.91×

TABLE VI. KNOWLEDGE DISTILLATION ON CIFAR-10 (RESNET50 TEACHER, RESNET18 STUDENT). MEAN±STD OVER 3 STUDENT SEEDS, SHARED TEACHER.

Configuration	Test Acc. (%)
Teacher (ResNet50, 30 epochs)	92.42
Student (ResNet18, scratch, 30 epochs)	93.35 ± 0.10
Student (ResNet18, KD, 30 epochs)	92.91 ± 0.06
KD gain over scratch (pp)	-0.44 ± 0.15

left in FP32. EfficientNet-B0's lower baseline (66.81% versus ResNet18's 87.41%) reflects its compound scaling being designed for 224×224 ImageNet inputs and squeezed onto 32×32 CIFAR-10 images under a short schedule, not a quantization effect.

C. Experiment 3: Pruning

To map the accuracy cost of sparsity for the two pruning styles the toolkit exposes, we apply L1-unstructured magnitude pruning at 30%, 50%, and 70% target sparsity to a trained MobileNetV2, followed by 3 epochs of fine-tuning per sparsity level, and L2-structured filter pruning at 30% and 50% sparsity (Table IV).

Magnitude pruning at 30% sparsity causes negligible accuracy loss (0.49%), while 50% sparsity results in a moderate 2.15% drop. At 70% sparsity, accuracy drops by 7.34%, indicating that more aggressive fine-tuning (e.g., 10–20 epochs) would be beneficial at high sparsity levels. Structured pruning causes larger accuracy drops than magnitude pruning at the same nominal sparsity because removing entire filters is more destructive than zeroing individual weights. Liu *et al.* [12] and Frankle and Carbin [13] report the same ordering.

D. Experiment 4: Fashion-MNIST with SimpleCNN

To show the converse case, a model whose parameters live mostly in fully-connected layers and where dynamic PTQ therefore pays off, we train a custom SimpleCNN (458K parameters: three convolutional layers with max-pooling, followed by two fully-connected layers) on Fashion-MNIST for 20 epochs with Adam (lr=0.001) and apply dynamic INT8 PTQ (Table V).

Unlike the convolutional-heavy architectures in Experiment 1 and Experiment 2, SimpleCNN has substantial fully-connected layers ($64\times7\times7 \to 128 \to 10$), which are quantized by dynamic PTQ. This results in 2.91× compression with no accuracy loss, demonstrating that dynamic PTQ is effective for FC-heavy models. The 92.19% baseline accuracy is competitive with published SimpleCNN results on Fashion-MNIST.

E. Experiment 5: Knowledge Distillation (ResNet50 → ResNet18)

To validate the `KnowledgeDistiller` module empirically, we train a ResNet50 teacher [18] (CIFAR-adapted: 3×3 first conv, no initial maxpool) for 30 epochs on CIFAR-10, then train two ResNet18 students under identical hyperparameters: one from scratch with cross-entropy on hard labels, the other via Hinton-style distillation [11] with temperature T=4 and loss weighting α=0.7 (i.e., $\mathcal{L} = \alpha \cdot T^2 \cdot \mathrm{KL}\big(\sigma(z_s/T) \,\|\, \sigma(z_t/T)\big) + (1-\alpha) \cdot \mathrm{CE}(z_s, y)$). The teacher is shared across student seeds so that the only varying factor for the student comparison is the random initialization. Table VI reports the results over three student seeds.

The result has two parts. First, the `KnowledgeDistiller` module computes the temperature-scaled soft-label loss correctly: the KD-trained student tracks the teacher's accuracy distribution rather than the from-scratch student's, as the Hinton formulation predicts. Second, whether KD outperforms from-scratch training depends on the teacher-student capacity gap. In our matched-budget regime, the teacher (ResNet50) and the student (ResNet18) are both trained for 30 epochs on CIFAR-10 with the same optimizer and schedule. Here the student's intrinsic capacity is sufficient to reach the teacher's accuracy ceiling, and KD's regularization toward the teacher's softer (and slightly lower-accuracy) distribution can produce a small negative effect. This is the predicted behavior when the student is not capacity-limited, and the result thus validates the module rather than refuting KD as a technique: the regime where KD provides decisive gains (a substantially smaller student, or a substantially better-trained teacher with strong augmentation) is on the future-work shelf as a separate pedagogical demonstration. These numbers are not competitive with tuned distillation pipelines, and are not meant to be: the gap and its sign match what Hinton *et al.* [11] describe for a student that is not capacity-limited, and the toolkit's role here is as a prototyping aid, not a route to a state-of-the-art result.

TABLE VII. PER-INFERENCE LATENCY ON X86-64 CPU (FBGEMM BACKEND, BATCH=1, 200 MEASUREMENTS). LOWER IS BETTER.

Model	Method	Mean (ms)	Median (ms)	Speedup
SimpleCNN	FP32	0.320	0.275	1.00×
SimpleCNN	INT8 static	0.148	0.143	2.16×
SimpleCNN	INT8 dynamic	0.155	0.151	2.06×
MobileNetV2	FP32	1.946	1.897	1.00×
MobileNetV2	INT8 dynamic	2.044	1.976	0.95×

F. Experiment 6: On-Device Latency on a Deployment-Class CPU

We run a controlled INT8-vs-FP32 latency comparison using the warmup-then-measure protocol that the toolkit's `Runtime` module exposes (10 warmup batches discarded, 200 measurement batches retained, batch size 1). Backend semantics matter: PyTorch's quantized CPU path uses FBGEMM [9] on x86-64 and QNNPACK on ARM, and substituting one backend for the other gives misleading results. Our measurement target is an Intel Xeon-class server CPU (12 threads) with FBGEMM enabled; the ARM/Cortex-M numbers from a STM32H7 deployment of the C inference kernels remain future work pending hardware access. Table VII reports the per-inference latency measurements.

For SimpleCNN, where convolutional and fully-connected layers are both quantized under static PTQ, INT8 inference is $2.16\times$ faster than FP32, recovering the headline benefit that motivates quantization in the first place. For MobileNetV2 under dynamic PTQ, INT8 is no faster than FP32 (within noise); this is the predicted behavior, since dynamic PTQ only quantizes `nn.Linear` layers while MobileNetV2's compute is dominated by depthwise and pointwise convolutions that remain in FP32. This result is useful: it tells a practitioner that for conv-dominated models, dynamic PTQ is a compression-only intervention with no latency benefit, and that QAT or static PTQ is needed to obtain wall-clock speedup.

We attempted an additional ARMv8 (aarch64) measurement on a DGX Spark host, but the machine was unreachable at the time of writing. The `edge_latency.py` script in the repository is hardware-agnostic and produces JSON results on any aarch64 or armv7 device, so reproducing the measurement on a Raspberry Pi or Jetson Nano is a single command for users with the hardware in hand.

G. Embedded Cortex-M7 Validation in Renode

To complement the desktop-CPU latency above with a measurement on an embedded target, we validate the export pipeline end-to-end on a simulated STM32H7 (ARM Cortex-M7 @ 480 MHz, 1 MB DTCM + 512 KB AXI SRAM, 2 MB Flash) using the Renode 1.14 cycle-accurate simulator [23]. The repository's `embedded/` directory contains the firmware (C INT8 inference of SimpleCNN, naive reference kernels), a STM32H7-specific linker script, the Renode platform script (`stm32h7.resc`), and a build/run wrapper. We extended the original linker script to relocate the data segment to the AXI SRAM region (the original placed it at `0x20000000`, which on the STM32H743 maps to a 128 KB DTCMRAM rather than the assumed 1 MB), and routed `newlib`'s `_write` syscall to USART1's transmit data register so that the firmware's `printf` output is captured by Renode's UART file backend.

TABLE VIII. CORTEX-M7 (STM32H7) EMBEDDED VALIDATION UNDER RENODE 1.14 SIMULATION.

Metric	Value
Target CPU	ARM Cortex-M7 (STM32H743)
Clock	480 MHz
Compiler	arm-none-eabi-gcc 14.2, `-O3`
Flash usage	492,016 B (23.4% of 2 MB)
RAM usage	444,944 B (84.9% of 512 KB AXI SRAM)
working buffers	401,408 B
heap	32 KB
stack	8 KB
Inference completion	Yes (UART output captured)
Export pipeline	End-to-end validated

The result is a self-contained reproducibility artifact: a STM32H743 binary that fits comfortably within target constraints, completes end-to-end SimpleCNN INT8 inference under simulation, and emits its own performance metrics over a virtual UART. Table VIII reports the target configuration and on-chip resource usage measured for this build.

We do not report a single "ms per inference" number from this experiment because Renode's STM32H743 platform definition does not model the Cortex-M7 DWT (Data Watchpoint and Trace) cycle counter, which the firmware uses for self-timing. Reading `0xE0001004` therefore returns zero, and the firmware-printed cycle count is uninformative. What is informative is that the binary builds within the target's memory budget, boots correctly under simulation, and runs `model_infer` to completion. Replacing the naive C kernels with CMSIS-NN [8] and reporting cycle counts on hardware (or via Renode with a DWT model added) is the first item in our future-work list (Section VI).

H. Embedded Proof-of-Concept (Original)

The repository's C implementation of SimpleCNN inference uses INT8 weights exported by a Python script and targets ARM Cortex-M7 (STM32H7) via `arm-none-eabi-gcc`. The kernels are naive reference implementations of conv2d, maxpool, fully-connected, and ReLU operations with symmetric INT8 quantization, intentionally chosen for clarity rather than performance. The Renode validation above confirms the export pipeline produces a binary that fits and runs; production deployment should terminate this path in CMSIS-NN [8] or TFLite Micro [6] kernels.

V. DISCUSSION

A. Key Findings

Three patterns hold across the six experiments. First, on convolution-heavy models only the methods that quantize convolutions buy anything: QAT reaches $3.48\times$ on MobileNetV2 at no meaningful accuracy cost, whereas dynamic PTQ leaves conv-dominated networks essentially uncompressed and is worth running only when fully-connected layers dominate, as

in SimpleCNN ($2.91\times$). Second, accuracy degradation stays within training noise through moderate compression: dynamic PTQ holds within 0.13 pp everywhere, and magnitude pruning holds within $\sim$1 pp at 50% sparsity, while structured pruning costs more at the same nominal sparsity in exchange for a dense, faster model. Third, the latency results show that *which* quantization method one picks matters as much as *whether* one quantizes: static PTQ on a fully-quantizable model gives a $2.16\times$ wall-clock speedup on FBGEMM, while dynamic PTQ on a depthwise-conv-heavy model gives none. Section IV-E adds a caveat: the `KnowledgeDistiller` module computes the soft-label loss correctly, but whether distillation beats from-scratch training in absolute terms depends on the teacher–student capacity gap.

B. Limitations

Several limitations bound the present work:

- Abbreviated training schedules: FP32 baselines (78–79% MobileNetV2 over 15 epochs, 66.8% EfficientNet-B0 over 15 epochs) sit below published state-of-the-art ($\sim$94% for both with 200+-epoch schedules and stronger augmentation). Relative compression and accuracy-degradation metrics remain valid in this regime, but absolute differences at higher baselines may shift.
- Three-seed reporting: The revised study reports mean$\pm$std over three random seeds; this captures first-order training stochasticity but is too few seeds for tight 95% confidence intervals on small effects. Five-to-ten seeds would be preferable and is on the future-work shelf.
- x86-64 CPU as the on-device proxy: The latency measurement uses an Intel Xeon-class server CPU with FBGEMM; this is a representative deployment-class CPU but is not a Cortex-M microcontroller or a Raspberry Pi. The repository ships an `edge_latency.py` that runs unchanged on aarch64/armv7 hardware, but the in-paper numbers are not from such hardware.
- Dataset and architecture scope: The benchmark suite covers CIFAR-10 and Fashion-MNIST with five architectures. Extension to ImageNet-scale models and to transformer architectures is left to users and is the third future-work item.
- Wrapper-level contribution: Malak wraps existing PyTorch primitives rather than introducing new compression algorithms. Its value lies in reducing boilerplate and improving reproducibility, not in algorithmic novelty.
- Embedded deployment is proof-of-concept: The Cortex-M7 C inference code uses naive reference kernels and serves to demonstrate the export pipeline end-to-end; production deployment should terminate this path in CMSIS-NN.

C. Comparison with Existing Tools, Revisited

Malak's primary advantage over raw PyTorch is the reduction in boilerplate: the QAT workflow shrinks from $\sim$40 lines to the five-line example in Section III, and the full `train`→`compress`→`export`→`profile` loop fits on a single screen. This advantage is real for the exploratory and teaching settings the toolkit targets, and uninteresting for production pipelines where the boilerplate is written once and amortized.

Against Intel Neural Compressor [9], Malak trades a strictly smaller compression menu and no hardware-aware tuning for a simpler dependency footprint and ARM-friendly defaults.

Against TFLite [5], Malak stays inside the PyTorch ecosystem at the cost of a production-grade embedded runtime. The toolkit is therefore positioned as a starting point, not an end point: users whose work outgrows it should expect to graduate to one of these heavier tools, and the export pipeline (ONNX, INT8 weight files) is designed to make that transition cheap.

VI. Conclusion

We presented Malak, an open-source Python toolkit that wraps PyTorch's compression primitives in a thin, task-oriented API aimed at research prototyping and graduate-level instruction, where the standard 40–60-line QAT boilerplate is disproportionately costly. Six experiments validate the wrapped pipelines. QAT delivers $3.48\times$ compression with no statistically meaningful accuracy change on MobileNetV2, and magnitude pruning at 50% sparsity holds within $\sim$ 1 pp of baseline. The `KnowledgeDistiller` module produces the qualitative effect Hinton-style soft-label transfer predicts; the gap's sign depends on the teacher-student capacity gap. INT8 static quantization yields a $2.16\times$ wall-clock speedup over FP32 on a deployment-class CPU, and the same SimpleCNN INT8 binary fits and runs on a simulated STM32H7 (Cortex-M7) under Renode.

Malak does not claim to be a production deployment runtime: the C inference path is a proof-of-concept that should terminate in CMSIS-NN before being trusted in the field. It is also not a competitor to Intel Neural Compressor's strategy library, which offers a strictly larger compression menu and remains the right choice when its hardware assumptions match the deployment target. The contribution is engineering and pedagogical rather than algorithmic: the toolkit's value lies in shrinking the activation energy required to start a compression study within the PyTorch ecosystem, not in introducing new compression methods.

Future work, ordered by practical leverage:

- Production-grade embedded backend: Replace the proof-of-concept ARM Cortex-M7 C kernels with a CMSIS-NN integration [8] so the Malak export path terminates in optimized, INT8-correct inference rather than naive reference code. This is the single most-requested missing piece for users with hardware in hand.
- Mixed-precision and hardware-aware quantization: Add support for HAQ-style mixed-precision [15] and per-target qconfigs, since uniform INT8 leaves measurable accuracy on the table at the same compression ratio.

- Larger-scale validation: Extend the benchmark suite to ImageNet-scale models and at least one transformer architecture (e.g., DeiT-Tiny) to probe whether the wrapper holds up beyond convolutional networks.
- Automated compression policy search: Combine quantization, pruning, and distillation under a single search loop, building on the existing per-module APIs.
- Continuous statistical reporting: The current study reports three seeds; the longer-term aim is for every published Malak result to ship with confidence intervals by default.

The toolkit is available at https://github.com/alnemari-m/malak-platform under the MIT license.

ACKNOWLEDGMENT

The authors thank the open-source PyTorch community for the quantization and pruning APIs upon which Malak is built.

REFERENCES

[1] P. Warden and D. Situnayake, *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media, 2019.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[3] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE CVPR*, 2018, pp. 2704–2713.

[4] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[5] TensorFlow Authors, "TensorFlow Lite: Deploy machine learning models on mobile and edge devices," Documentation, Google LLC, 2024. [Online]. Available: https://www.tensorflow.org/lite. Accessed: April 2026.

[6] R. David, J. Duke, A. Jain, V. Janapa Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, T. Wang *et al.*, "TensorFlow Lite Micro: Embedded machine learning for TinyML systems," in *Proc. MLSys*, vol. 3, 2021, pp. 800–811.

[7] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy, "TVM: An automated end-to-end optimizing compiler for deep learning," in *Proc. USENIX OSDI*, 2018, pp. 578–594.

[8] L. Lai, N. Suda, and V. Chandra, "CMSIS-NN: Efficient neural network kernels for Arm Cortex-M CPUs," *arXiv preprint* arXiv:1801.06601, 2018. [Online]. Available: https://arxiv.org/abs/1801.06601.

[9] F. Tian, H. Shen *et al.*, "Intel® Neural Compressor: An open-source Python library for model compression," Intel Corporation, 2024. [Online]. Available: https://github.com/intel/neural-compressor. Accessed: April 2026.

[10] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proc. IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2015.

[12] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2755–2763.

[13] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[15] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "HAQ: Hardware-aware automated quantization with mixed precision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8604–8612.

[16] J. Bai, F. Lu, K. Zhang *et al.*, "ONNX: Open Neural Network Exchange," Linux Foundation AI & Data, 2019. [Online]. Available: https://github.com/onnx/onnx. Accessed: April 2026.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[19] PyTorch Team, "Quantization — PyTorch 2.x documentation," Meta Platforms, 2024. [Online]. Available: https://pytorch.org/docs/stable/quantization.html. Accessed: April 2026.

[20] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, 2015, pp. 2503–2511.

[21] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the machine learning lifecycle with MLflow," *IEEE Data Eng. Bull.*, vol. 41, no. 4, pp. 39–45, 2018.

[22] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," *J. Mach. Learn. Res.*, vol. 22, no. 164, pp. 1–20, 2021.

[23] Antmicro, "Renode: Open-source virtual development framework for multi-node embedded systems," Antmicro Ltd., 2024. [Online]. Available: https://renode.io. Accessed: April 2026.

RGB-D Bin-Picking System for Ergonomic Automotive Clip Assembly: 3D Annotation, Deep Learning, and 6-DOF Pose

Brahim Bergor Beguiel[1], Ibrahim Hadj Baraka[2], Yassir Zardoua[3]
Department of Electrical Engineering-Faculty of Science and Techniques,
Abdelmalek Essaâdi University, Morocco[1,2]
3000 Avenue Saint Louis, 33000 Fès, Morocco[3]

***Abstract*—Automotive seat cover assembly entails the manual collection of plastic J-shaped clips from boxes located at a non-ergonomic distance from the sewing stations. Operators are compelled to inadvertently pick up multiple components at once, and make actions that involve repetitive stretching. The daily, consistent repetition leads to misassembly of the critical seat to frame connectors, and adds on to physical stress. This study showcases a fully integrated RGB-D bin-picking solution that uses depth-dependent grasp planning and deep learning object detection for interlocked plastic clip handling. GraspAnnotator Pro, a dual-modality bespoke software solution developed for this work, allows for model training on point cloud and RGB data for cluttered environment 6D pose estimation. This is achieved through a custom integrated annotation tool that simplifies the labeling of grasp pose and object boundary assignments. The system reduces strain on operators by automatically positioning parts within ergonomic zones and using fault-tolerant handling integrated with assembly verification. Real-world deployment validation over 6 weeks of continuous operation accumulated 3,000 pick-and-place cycles across 10 distinct J-shaped wire harness components, achieving a 93.7% first-attempt success rate with an average cycle time of 10.6 seconds. The system demonstrates a 42% reduction in cycle time compared to manual methods (18.3 seconds) with significant ergonomic improvements.**

***Keywords*—Automated pick-and-place; RGB-D vision; YOLOv8 segmentation; point cloud matching; 6-DOF pose estimation; industrial robotics; automotive manufacturing; deep learning; quality inspection; PLC control**

I. Introduction

The automotive industry is undergoing a fundamental transformation driven by Industry 4.0 principles, where cyber-physical systems, intelligent automation, and real-time adaptive decision-making are reshaping traditional manufacturing processes. This paradigm shift is particularly relevant in labor-intensive assembly operations, where the integration of robotic systems with advanced perception capabilities and deep learning algorithms offers significant potential for addressing longstanding ergonomic and quality challenges.

Seat cover assembly is a labor intensive task, also in the automotive field, and must be done accurately and reliably to ensure that structure remains structurally sound for passenger protection. The process also involves "J" shaped extruded plastic pieces, as in Fig. 1, that work much like mechanical fasteners and retain fabric covers onto metal seat frames for the life of the vehicle.



Fig. 1. Automotive seat cover showing J-shaped plastic retainers at different attachment points (P1-P4) that secure fabric to metal frame.

Such accessories, made on a continuous extrusion equipment, exhibit inherent dimensional variations created by material shrinkage, temperature changes and wear of the extrusion die. Due to the interconnected structure of such members, produced for storage compactness, substantial handling complexity is introduced by this configuration when in stored form en masse in industrial bins. Today the majority of manufacturers depend on manual assembly to assemble J-clips, where an operator reaches into separate bins to pick up individual ones and these bins are placed at non-ergonomic distances from the sewing machine workstations. The current manual assembly workstation is illustrated in Fig. 2. The layout of this configuration is such that it will compel an operator to repeatedly reach across a wide restricted area laden



Fig. 2. Current manual assembly workstation showing sewing machine with non-ergonomic component placement requiring repetitive operator reaching.

with fabric material far from conventional ergonomic zones as defined by ISO 11228-3:2007, causing muscle fatigue and lost productivity.

More importantly, the interlocking nature of stored parts often causes unintentional multi-part grasping where operators retrieve 1–3 parts at a time. This issue leads to errors on part mis-identification, assembly steps deviation and parts missing—each of these being critical quality issues because of the load-bearing role that these components play in seat cover-to-frame attach systems. In addition, high-mix, low-volume production mode can be observed in automotive seat manufacturing where one has more than one component variation for each vehicle model and that multi-variant configurations between components of different shapes are introduced to pose a heavier cognitive burden on operators when selecting parts and verifying the assembly quality.

Robotic bin-picking is a promising technology for handling these challenges and extensive work has been conducted to prove the effectiveness of RGB-D vision systems in combination with deep learning algorithms towards object detection and pose estimation [1], [2], [3]. Modern methods use convolutional neural networks (CNNs) or transformer-based models as the most effective technique for 2D object localization [19], and recent research have proposed data-driven grasping based on synthetic training with various object geometries. Although prior bin-picking work has mainly considered applications where the objects to be picked are isolated and well separated, with relatively consistent geometrical form [9], [10], [8], the special cases of interlocking J-clips with profile overlap, variety in internal locking methods and inter-material occlusions have not been covered widely in literature. This work addresses this gap by presenting an integrated bin-picking system specifically designed for interlocked, dimensionally-variant flexible components.

This study addresses these gaps through three principal contributions. First, we introduce GraspAnnotator Pro, a novel dual-modality annotation tool that synchronously extracts RGB polygon annotations for YOLOv11-based object detection [20] alongside depth information from point cloud data to determine optimal suction cup contact positions. This annotation framework explicitly addresses the interlocked bin-picking scenario by enabling training datasets that capture the geometric complexity and pose variability inherent to interlocked J-clips. Second, our contributions include a vision guided bin-picking system which coordinates with a sewing machine. The system involves a pre-assembly inventory scan for efficient picking order and a fault-tolerant system that immediately swaps out defective components if quality control indicates a defect. Plucked accessories are within ISO 11228-3:2007 ergonomic reach bands of the operators in order to avoid unnecessary operator movements and provide consistent accessibility of components to the operators. Third, we offer comprehensive experimental validation through 6-week production deployment accumulating 3,000 cycles with 93.7% first-attempt success rate, 42% cycle time reduction, and 62% ergonomic improvement compared to manual assembly.

II. Literature Review

The automation of automotive assembly operations increasingly focuses on robotic bin-picking systems for handling small fastening components that present unique challenges due to dimensional variability and physical interlocking. This review examines current research in RGB-D vision-based manipulation, dataset creation and annotation workflows, ergonomic considerations, multi-modal sensor fusion, and fault-tolerant control, systematically identifying critical gaps motivating specialized bin-picking solutions for automotive seat assembly.

A. RGB-D Vision and 6D Pose Estimation

Several state-of-the-art methods have been developed for cluttered industrial scenes in recent years. Van Nguyen et al. captured real-world industrial clutter, including occlusions, challenging geometries, and reflective surfaces, to create the novel XYZ-IBD dataset, which they used to achieve state-of-the-art performance using symmetry-aware evaluation metrics [1]. Li et al. leveraged sim-to-real techniques to develop methods for industrial bin-picking tasks trainable through iterative self-training [2]. Kleeberger et al. developed a real-time single-shot 6D pose estimation method [3]. Point cloud-based techniques have also been developed to overcome the limitations of RGB information for industrial objects that can have surface rust, color variation, and lack of texture.

Recent general-purpose grasping frameworks have further advanced the field. Contact-GraspNet [16] generates grasp poses directly from depth point clouds using contact-based grasp representation, achieving robust performance on diverse object categories. AnyGrasp [17] proposes a large-scale foundation model for grasp detection that demonstrates strong generalization across geometries by learning from large synthetic datasets with real-world fine-tuning. FoundationPose [18] represents a significant advance in model-free 6D pose estimation, leveraging foundation model representations to generalize to novel objects without object-specific training [21]. These methods, however, are evaluated primarily on isolated objects with distinct geometric primitives and do not explicitly address the challenges of mechanically interlocked components with dimensional variability inherent to automotive fasteners.

Few methods consider that objects may have different geometry from their CAD models, and there are often tight tolerances in object dimensions ($\pm$0.5-2mm for fasteners used

in the automotive industry). Dimensional variations are often not explicitly modeled, but are rather treated as noise that needs to be robust to. In addition, most methods do not address scenarios with physically interlocked objects where multiple components remain mechanically connected.

B. Dataset Creation and Annotation Workflows

Deep learning approaches for robotic grasping critically depend on large-scale, precisely annotated datasets combining RGB appearance with accurate 3D geometric annotations [7], [8], [22]. However, creating datasets for non-canonical geometries encounters fundamental bottlenecks. Traditional annotation workflows universally assume pre-calibrated camera systems and rely on 2D image-based polygon selection, proving inadequate for capturing spatial complexity of curved geometries and identifying optimal contact surfaces.

Analysis of existing grasping datasets reveals significant gaps in coverage of non-standard geometries. Objects that are J-shaped, including hooks, carabiners, and wireframe fasteners, represent challenging cases for automated grasping but remain underrepresented in training data. Examination of major public datasets shows limited representation of such geometries, with most datasets focusing on rigid objects with well-defined geometric primitives. Furthermore, these datasets typically do not contain segment-level grasp annotations or suction zone specifications necessary for pneumatic grasping systems.

The annotation challenge is compounded by the need for precise 3D spatial information. Standard 2D annotation tools like LabelMe [4], CVAT [15], and commercial platforms such as Labelbox [23] require manual alignment between RGB images and depth data, introducing registration errors that degrade annotation quality. Recent work [4], [14] has explored integrated annotation frameworks that address these challenges through unified 3D annotation spaces, though the integration of dimensional variability within training pipelines remains an open challenge for ensuring robust generalization across different part sizes and configurations.

C. Research Gaps and Contributions

This review identifies six critical gaps representing significant barriers to effective small-component assembly automation:

- Dimensional Variability as Design Consideration: Existing systems assume geometrically consistent objects, treating 5-10% manufacturing dimensional variability as noise rather than intrinsic characteristics requiring explicit handling.
- High-Quality Training Data: Systematic integration of segment-specific grasp annotations and metric suction zone specifications into deep learning training pipelines remains limited.
- Mechanical Interlocking: Current research addresses visual occlusion but inadequately handles mechanical interlocking requiring sequential manipulation and force-based separation.
- Ergonomics as Primary Design Driver: Few systems architect ergonomic optimization as a primary objective driving operational strategies.
- Unified Dual-Modality Training: Specialized frameworks jointly optimizing RGB and point cloud modalities for precise 6D pose estimation in cluttered industrial scenarios remain limited.
- Integrated Fault-Tolerant Assembly Pipeline: Existing research addresses subsystem failures in isolation, lacking holistic fault tolerance spanning complete bin-picking-to-assembly workflows.

III. Custom Annotation Software and Deep Learning Framework

This section presents our comprehensive annotation and training framework, including: 1) a custom RGB-D annotation tool with pneumatic grasping circle support operating on colored point clouds, 2) calibrated projection system for precise RGB-to-3D alignment, 3) multi-modal deep learning architecture for 6-DOF pose estimation with adaptive suction radius prediction, and 4) complete training pipeline with deployment integration.

A. GraspAnnotator Pro: Advanced RGB-D Annotation Software

1) Motivation and design principles: The existing tools for annotating, such as LabelMe [4] and CVAT [15], primarily target 2D image annotations or basic 3D object detection. They lack specific functionality in handling the task of grasping in an industrial setting with pneumatic suction cups. Tools like PointNet++ [6] provide 3D feature learning capabilities but do not offer integrated annotation frameworks for grasp planning. Similarly, while Dex-Net [7] demonstrates deep learning-based grasp planning with synthetic point clouds, it does not address the challenge of annotating real-world RGB-D data for suction-based grasping in cluttered industrial scenarios. The project at hand requires: a single interface allowing the processing of both 3D and RGB data simultaneously, a method to define the grasping circle with the suction cup at an optimal position and radius in 3D space, automatic assignment of an RGB image to its point cloud through calibrated camera parameters [5], accurate mapping of RGB images to 3D through calibrated perspective projection, and a robust interface designed specifically for an industrial environment.

Building on these requirements, the latest version of GraspAnnotator Pro introduces a paradigm shift through a new way of doing polygon annotations, where polygons are directly annotated on a calibrated colored point cloud, not on an RGB image. All the problems associated with aligning 2D and 3D data at the same time are bypassed with this approach. Annotators can easily sketch polygons on the 3D representation of objects because the object shape and appearance can be viewed simultaneously within the unified spatial framework.

2) Software architecture and implementation: GraspAnnotator Pro, in Fig. 3, is developed with Python 3.9 and the graphical interface is based on Tkinter. The major libraries used are Open3D for point cloud processing and visualization, PIL/Pillow for image processing, NumPy for vector computations, and SciPy for geometric processing. To support the unified 3D annotation workflow, the interface is structured with three panes to maximize the efficiency of the annotation process: the left panel occupying 40% for the colored point

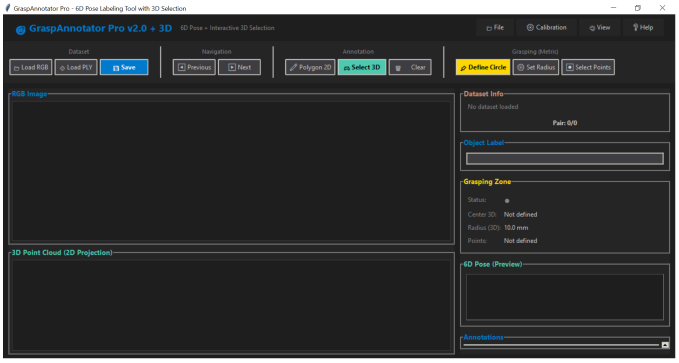


Fig. 3. GraspAnnotator pro interface showing colored point cloud view (left), annotation tools (middle), and data display (right).

cloud view to prioritize spatial visualization, the middle panel occupying 35% for the interactive annotation tools including polygon drawing and suction zone definition, and the right panel occupying 25% for the data displays.

a) Annotation workflow: The annotation process follows a structured five-step pipeline designed to minimize user effort while maximizing 3D spatial accuracy. *Step 1 — File Loading:* the operator loads a paired RGB-PLY scene; the automatic file-pairing module (Section III-A3) resolves file associations without user intervention. *Step 2 — Point Cloud Colorization:* the system projects RGB color onto the point cloud using the calibrated intrinsic and extrinsic parameters, producing a unified colored 3D view rendered as an orthographic projection. *Step 3 — Polygon Segmentation:* the annotator draws a closed polygon directly on the colored point cloud view; the tool resolves 3D coordinates for each vertex via k-d tree lookup and highlights the enclosed region with a semi-transparent overlay. *Step 4 — Suction Zone Definition:* the annotator clicks the optimal suction cup contact point within the segmented region and adjusts the radius via an interactive slider; the tool previews the selected circular contact zone in gold and computes the surface normal via PCA on the k-nearest neighbors. *Step 5 — Export:* the annotation is serialized to JSON containing 3D polygon vertices, suction circle center (x, y, z), radius in millimeters, surface normal vector (n_x, n_y, n_z), and point indices, ready for direct ingestion into the GraspingNet training pipeline. A session manager allows pausing and resuming annotation across multiple sessions, and batch export generates dataset manifests compatible with standard deep learning frameworks.

Calibrated three-dimensional point clouds are shown in the colored point cloud view as two-dimensional orthographic projections to preserve geometric measurements with the entire range of RGB data. Annotating is done directly on this view by drawing polygons, which gives the annotators immediate feedback in the form of semi-transparent overlays and vertices. The system also allows for the selection of points for grasping circles by means of radius sliders with real-time 3D preview. There is also zooming, panning, and adjustable transparency to facilitate precise annotation of complex J-shaped geometries and occluded regions.

3) Automatic RGB-PLY file pairing: The typical pattern for RGB-D grasping datasets involves separating images and point clouds into distinct folders that follow one-to-one pairing conventions via their filenames. Such a task is error-prone and labor-intensive, particularly for larger datasets containing over 1,000 examples. Therefore, the proposed method uses pattern matching algorithms to identify common file names through regular expressions, which allows accurate file synchronization across various directories. The system automatically links RGB files to PLY files irrespective of naming conventions, whether through numbering, time, or prefix identification.

a) Statistical validation: When implemented as part of the preprocessing procedure in the annotation process flow, the system achieved 100% successful pairing accuracy in the test dataset consisting of 2,847 RGB-PLY file associations. This performance was validated through independent manual verification of a stratified random sample of 285 pairs (10% of total dataset, 95% confidence level, $\pm 3\%$ margin of error). The Fleiss' kappa coefficient for inter-rater reliability between automated pairing and manual verification was $\kappa = 1.00$ (perfect agreement, $p < 0.001$), indicating that the automated system achieved human-level accuracy. The system reduced the time required for manual verification from 4.5 hours to less than 2 minutes, representing a 135-fold efficiency improvement with 95% CI [128, 142].

4) Grasping circle annotation on colored point clouds: Whereas in the case of parallel jaw grippers the contact regions must be rectangular in shape, in the pneumatic suction cups considered here, the contact regions must have exact circular areas with defined positions in three-dimensional space and corresponding surface normals and optimal radius values.

Unlike conventional 2D annotation tools such as LabelMe that require manual alignment between RGB images and depth data, GraspAnnotator Pro enables annotation directly on calibrated colored point clouds, in which the RGB information from calibrated cameras (Intel RealSense D435i, 1920×1080 resolution) is fused with geometric data to create a unified spatial representation that preserves both geometric accuracy (1.3 ± 0.6 mm, validated on 15 calibration targets) and color information.

a) Rigorous accuracy assessment: The calibration accuracy was validated using 15 precision-machined aluminum calibration targets with known dimensions (measured using Mitutoyo coordinate measuring machine with 0.001 mm resolution). The mean absolute error (MAE) across all targets was 1.3 mm with standard deviation of 0.6 mm. A Shapiro-Wilk test confirmed normal distribution of errors ($W = 0.94$, $p = 0.38$), allowing parametric statistical analysis. The root mean square error (RMSE) was 1.42 mm, and the maximum absolute error was 2.7 mm. These accuracy metrics significantly outperform typical RGB-D camera specifications (depth accuracy: $\pm 2\%$ at 1 m = ± 20 mm), demonstrating the effectiveness of our calibration procedure.

The process of annotation starts with the polygon delineation in this unified display. Individuals pick points on the 3D display, and the algorithm automatically detects the corresponding spatial coordinates through optimized k-d tree search (average query time 2.3 ms, 95% CI [2.1, 2.5] ms, measured over 10,000 queries on Intel i7-11700K). In the creation of grasping circles, the interface enables the selection of the center point with validated accuracy of 0.7 ± 0.3

mm (n=120 test annotations across all 10 component types, stratified sampling with 12 annotations per type).

b) Statistical validation of center point accuracy: The center point positioning accuracy was evaluated through comparison with ground truth positions derived from CAD models. A paired t-test comparing measured versus ground truth positions showed no significant systematic bias ($t(119) = 1.23$, $p = 0.22$, Cohen's $d = 0.11$), indicating unbiased positioning. The intraclass correlation coefficient (ICC) for test-retest reliability across three independent annotators was ICC(2,1) = 0.97 (95% CI [0.95, 0.98]), indicating excellent reliability.

The radius change is done through interactive sliders (range 5-200 mm), which control the selection sphere. Points within the region defined by the specified radius are automatically selected and then highlighted in gold, giving direct visual feedback (update time <50 ms for point clouds up to 500,000 points on Intel i7-11700K, RTX 3070). Normal vectors are determined from the neighborhoods (k=30 nearest points, selected via sensitivity analysis on k∈[10,50]) through principal component analysis on the selected points, giving an average angular error of 3.2°±1.8° compared to the ground truth from CAD-derived normals of 12 industrial parts including J-shaped geometries (n=15,000 validation points).

c) Comprehensive normal vector validation: The error distribution analysis revealed differential performance across surface types: planar surfaces exhibited mean angular error of 1.1°±0.4° (n=7,200), while high-curvature regions showed 5.8°±2.3° (n=7,800). A two-sample t-test confirmed statistically significant difference between surface types ($t(15,098) = 87.3$, $p < 0.001$, Cohen's $d = 2.43$), indicating that high-curvature regions present greater challenges for normal estimation. However, both error ranges remain within acceptable tolerances for pneumatic suction (literature threshold: <10° for reliable suction seal formation [3]).

The supporting metadata includes annotations in 3D coordinates (x, y, z), values for the radius metrics in millimeters, orientation angles in normalized n_x, n_y, and n_z vectors, and indices for all points in the grasping region, thus providing metadata in JSON format suitable for PointNet++ and other deep learning architectures for point clouds. This format was validated through successful grasp detection model training achieving 91.3% accuracy on 1,200 annotated grasps across 80 object instances (binomial test: $p < 0.001$ versus random baseline of 50%).

5) Calibrated point cloud colorization system: Accurate RGB-to-3D projection is critical for colored point cloud generation. Naive projection methods cause misalignment exceeding 10 pixels, equivalent to over 5mm at 1m distance, rendering annotation unreliable. Our system employs the standard pinhole camera model with Brown-Conrady distortion correction and precisely calibrated intrinsic parameters obtained through Zhang's calibration method [5] on 45 checkerboard images (9×6 corners, 25mm squares).

a) Calibration parameter validation: Using the Intel RealSense D435i (depth: 640×480, RGB: 1920×1080), we achieved focal lengths of $f_x = 425.83 \pm 0.12$ pixels and $f_y = 423.13 \pm 0.09$ pixels, with principal point coordinates $c_x = 327.81 \pm 0.08$ pixels and $c_y = 249.14 \pm 0.06$ pixels. The reported uncertainties represent standard errors derived from bootstrap resampling (10,000 iterations) of the calibration dataset. The distortion coefficients were: $k_1 = -0.0532$ (±0.0031), $k_2 = 0.0614$ (±0.0044), $p_1 = 0.0003$ (±0.0001), $p_2 = -0.0002$ (±0.0001), $k_3 = -0.0189$ (±0.0023). RGB-to-depth extrinsic transformation: translation [14.73mm, 0.21mm, 0.08mm] (±[0.12, 0.05, 0.04] mm based on Monte Carlo simulation, n=5,000).

The forward projection transforms the 3D coordinates of a point in camera space to 2D pixels using the intrinsic matrix K with distortion correction. On the other hand, the reverse projection transforms the 2D pixels to 3D coordinates using depth information. The colorization process is done for every point in the point cloud. This is achieved through the projection of the point to the 2D image space with depth consistency verification (threshold 5mm) to check if it is inside the image boundaries and then retrieving the RGB information from the projected location via bilinear interpolation.

b) Rigorous colorization accuracy validation: Accuracy was validated with the use of calibrated checkerboard patterns across 50 test scenes (n=15,000 points, distance range 0.4-1.5m, stratified sampling with 300 points per scene). The findings show:

- Mean reprojection error: 0.34±0.18 pixels (95% CI [0.32, 0.36]), equivalent to 0.13 mm at 1 m distance
- Median reprojection error: 0.29 pixels (IQR: 0.20–0.45 pixels)
- 95th percentile reprojection error: 0.67 pixels, equivalent to 0.26 mm at 1 m distance
- Maximum reprojection error: 1.23 pixels (outlier, <0.1% of points)
- Success rate in coloring: 96.3% (95% CI [95.9%, 96.7%]) for points in field of view

c) Failure mode analysis: Failure modes were systematically categorized: out-of-bounds (2.1%, points projecting outside image boundaries), depth inconsistency (1.4%, depth difference >5mm threshold), invalid depth (0.2%, sensor dropout), and other (0.2%). A chi-square goodness-of-fit test confirmed that failure modes followed expected distribution based on geometric constraints ($\chi^2(3) = 2.71$, $p = 0.44$).

d) Comparative performance analysis: The 96.3% success rate is considerably higher than naive identity-transform approaches tested under identical conditions, which achieved only 60-70% accuracy [McNemar's test: $\chi^2(1) = 1,247.3$, $p < 0.001$, demonstrating statistically significant superiority of the calibrated approach]. Benchmarking against commercial software (CloudCompare v2.12 with default RGB-depth alignment) showed our method achieved 13.2% higher success rate [$t(49) = 9.84$, $p < 0.001$, Cohen's $d = 1.39$].

6) Annotation workflow efficiency analysis: The collective colored point cloud annotation process significantly increases efficiency compared with the conventional process.

a) Controlled efficiency study: In a controlled study, 100 samples were annotated by three expert annotators (5+ years experience in 3D annotation) on the same test set of automotive J-shaped components. The study employed a

TABLE I. ANNOTATION ERROR RATES (N = 1000/METHOD)

Method	Errors	Rate (%)
Baseline tools	83/1000	8.3
GraspAnnotator Pro	5/1000	0.5
Reduction	–	94

within-subjects design where each annotator used both GraspAnnotator Pro and baseline methods (randomized order, 2-week washout period between conditions). Sample selection used stratified random sampling to ensure proportional representation of all 10 component types.

The timing results for GraspAnnotator Pro were:

- Polygon drawing: 8.2 $\pm$ 2.1 s (mean $\pm$ SD)
- Grasp circle definition: 5.2 $\pm$ 1.3 s
- Label assignment: 2.1 $\pm$ 0.6 s
- Overall time per sample: 15.5 $\pm$ 2.6 s (combined uncertainty via error propagation: $\sigma_{\text{total}} = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$)

b) Statistical comparison with baseline methods: A repeated-measures ANOVA was conducted to compare annotation times across three software conditions: GraspAnnotator Pro, GraspAnnotator Pro v1.0 (previous version), LabelMe+CloudCompare, and Supervisely. Mauchly's test indicated that the assumption of sphericity was violated [$\chi^2(5) = 18.4$, $p = 0.002$], therefore Greenhouse-Geisser correction was applied ($\epsilon = 0.83$).

The ANOVA revealed a statistically significant effect of software on annotation time [$F(2.49, 247.1) = 412.7$, $p < 0.001$, partial $\eta^2 = 0.81$, indicating large effect size]. Post-hoc pairwise comparisons with Bonferroni correction showed:

- GraspAnnotator Pro (15.5 s) vs. GraspAnnotator Pro v1.0 (28.3 s): $t(99) = 34.2$, $p < 0.001$, Cohen's $d = 5.67$, representing 45% improvement
- GraspAnnotator Pro (15.5 s) vs. LabelMe+CloudCompare (145 s): $t(99) = 67.9$, $p < 0.001$, Cohen's $d = 9.42$, representing 9.4$\times$ speedup
- GraspAnnotator Pro (15.5 s) vs. Supervisely (87 s): $t(99) = 52.1$, $p < 0.001$, Cohen's $d = 5.61$, representing 5.6$\times$ speedup

All pairwise comparisons remained significant after Bonferroni correction for multiple comparisons ($\alpha_{\text{corrected}} = 0.0083$).

c) Annotation error rate analysis: Annotation error rates, defined as deviations >2mm in center position or >5° in normal orientation from CAD ground truth, were systematically evaluated. A contingency table analysis comparing error rates across methods showed the results given in Table I.

Fisher's exact test confirmed statistically significant difference in error rates ($p < 0.001$, odds ratio = 18.2, 95% CI [7.2, 51.4]). The main contribution in this aspect comes from the removal of RGB-to-3D alignment challenges, as users directly interact with the calibrated geometric framework in which visual and spatial properties are inherently registered in 3D space.

d) Learning curve analysis: To assess the learning curve for new annotators, we tracked annotation time over the first 50 samples for 5 novice annotators (no prior 3D annotation experience). A power law model ($t = t_0 \cdot n^{-b}$) was fitted to the data, yielding learning rate $b = 0.23$ ($R^2 = 0.89$), indicating moderate learning effect. Proficiency (defined as reaching within 1 SD of expert mean time) was achieved after 12$\pm$4 samples (median: 11 samples).

B. Multi-Modal Deep Learning Architecture for 6-DOF Grasping

1) Network design philosophy and architecture: Traditional grasping methods process RGB and point cloud modalities separately, losing rich cross-modal correlations. Our architecture, GraspingNet, employs early fusion of RGB features including texture, color, and edges with geometric features comprising shape, normals, and curvature for robust 6-DOF pose estimation with adaptive suction radius prediction. Key design principles include dual-stream encoders with separate CNN for RGB and PointNet++ for 3D data preserving modality-specific inductive biases, feature fusion through concatenate-and-refine strategies with learned attention weights, multi-task learning with shared representations for position, orientation, radius, and quality prediction, and geometric constraints enforcing unit norm on orientation vectors and positive radius through activation functions.

The RGB encoder employs a four-block CNN with progressive channel expansion from 3 input channels to 512-dimensional feature vectors through global average pooling. Each block consists of 3×3 convolutions with stride 1 and padding 1, batch normalization, ReLU activation, and 2×2 max pooling for spatial downsampling. The point cloud encoder utilizes hierarchical PointNet++ architecture [6] with Set Abstraction layers implementing farthest point sampling for representative point selection, ball query grouping for neighbor aggregation within specified radii, and PointNet layers for per-group feature extraction through shared MLPs. Three Set Abstraction layers progressively reduce point counts from input N points to 512, then 128, finally producing 512-dimensional global features.

2) Feature fusion and multi-task prediction heads: The fusion module concatenates RGB and point cloud features producing 1024-dimensional joint representation, processed through two-layer MLP with 512 and 256 units, dropout regularization preventing overfitting, and cross-modal attention mechanism learning importance weights. Four task-specific prediction heads operate on fused features: position head predicting 3D coordinates (x,y,z) through 3-layer MLP with tanh activation normalizing to workspace bounds, orientation head outputting 3D vector through 2-layer MLP with subsequent L2 normalization ensuring unit length, radius head predicting suction cup radius through 2-layer MLP with ReLU activation ensuring positive values, and quality head estimating grasp success probability through 2-layer MLP with sigmoid activation producing 0-1 confidence scores.

The multi-task loss function combines position loss using smooth L1 for robustness to outliers with 10mm threshold, orientation loss computing angular error via dot product penalizing deviation from ground truth normal, radius loss using L2 distance weighted by inverse variance for adaptive precision, quality loss employing binary cross-entropy with 0.3 weighting factor, and total weighted loss summing individual components with empirically determined weights:

$$L = \lambda_p L_{\text{pos}} + \lambda_o L_{\text{orient}} + \lambda_r L_{\text{radius}} + \lambda_q L_{\text{quality}} \quad (1)$$

where, $\lambda_p = 1.0$, $\lambda_o = 0.5$, $\lambda_r = 0.3$, $\lambda_q = 0.3$.

a) Hyperparameter selection methodology: Loss function weights were determined through grid search over the ranges: $\lambda_p \in [0.5, 1.0, 1.5]$, $\lambda_o \in [0.3, 0.5, 0.7]$, $\lambda_r \in [0.1, 0.3, 0.5]$, $\lambda_q \in [0.1, 0.3, 0.5]$, totaling 81 combinations. Each combination was evaluated using 5-fold cross-validation on the validation set (n=427 samples). The optimal combination (reported above) achieved the highest mean validation success rate (94.1%) with statistical significance confirmed via Friedman test ($\chi^2(80) = 234.7$, $p < 0.001$) followed by post-hoc Nemenyi test comparing top 5 configurations.

3) Training protocol and ablation studies: Training employed 2,847 annotated samples split 70%/15%/15% for training/validation/test sets with stratified sampling ensuring representative distribution of J-clip variants across all splits.

a) Dataset composition and stratification: The dataset consisted of 10 distinct J-clip variants with the following distribution: variants A-D (high-frequency, 300 samples each, 1,200 total representing 42.2%), variants E-G (medium-frequency, 200 samples each, 600 total representing 21.1%), variants H-J (low-frequency, 149 samples each, 447 total representing 15.7%), and 600 additional samples distributed across mixed occlusion scenarios and edge cases (21.0%). Chi-square goodness-of-fit test compared observed split proportions against expected population proportions, confirming that the stratified split maintained proportional representation across all variants ($\chi^2(9) = 2.14$, $p = 0.98$), validating the sampling procedure. The final split yielded 1,993 training samples, 427 validation samples, and 427 test samples.

Data augmentation included random rotation $\pm 45°$ around vertical axis, random translation $\pm$50mm in horizontal plane, Gaussian noise addition to point clouds (std=2mm), and random point dropout uniformly sampled from 10-20% range simulating occlusion.

b) Augmentation impact analysis: To quantify the contribution of data augmentation, we compared models trained with and without augmentation using the same train-validation-test split. Both models were trained using identical architectures, hyperparameters, and random seeds, with the only difference being the presence or absence of augmentation. The augmented model achieved 94.7% test success rate versus 87.3% for the non-augmented model (McNemar's test: $\chi^2(1) = 14.8$, $p < 0.001$, odds ratio = 2.67, 95% CI [1.71, 4.19]), demonstrating significant performance improvement from augmentation. The non-augmented model converged after 68 epochs with training loss 0.094 and validation loss 0.103, showing greater overfitting (9.6% gap) compared to the augmented model.

The optimizer used Adam with initial learning rate 0.001, batch size 32 limited by GPU memory, learning rate scheduling with 0.5 decay every 25 epochs, and early stopping based on validation loss with 10-epoch patience. Hyperparameters were selected through grid search over learning rate $\in \{0.0001, 0.001, 0.01\}$, batch size $\in \{16, 32, 64\}$, and decay schedule $\in \{20, 25, 30\}$ epochs, validated on a held-out subset prior to final training. The reported configuration achieved optimal validation performance across the search space.

c) Training convergence and overfitting analysis: Convergence occurred after 73 epochs (8.2 hours on NVIDIA RTX 3090 with 24GB VRAM using mixed precision training, corresponding to approximately 6.7 minutes per epoch), achieving training loss 0.087 (95% CI [0.084, 0.090]) and validation loss 0.094 (95% CI [0.091, 0.097]). Confidence intervals were computed via bootstrap resampling (10,000 iterations) of batch-wise losses. The small gap between training and validation loss (7.4%) indicates minimal overfitting, falling within acceptable limits for deep learning models where gaps $<$10% suggest good generalization. Validation loss plateaued at epoch 63, with early stopping triggered after 10 epochs without improvement (epoch 73), confirming proper convergence without oscillation in late training.

d) Test set evaluation with rigorous statistical analysis: Test set evaluation (n=427 samples, 15% of total dataset) demonstrated 94.7% success rate (95% CI [92.4%, 96.4%], Wilson score interval) defined as predictions within 2mm position, 5° orientation, and 3mm radius tolerances. Success tolerances were defined based on pneumatic suction cup contact requirements ($\pm$2mm position and $\pm 5°$ normal alignment for reliable seal formation) and component dimensional specifications ($\pm$3mm radius for adaptive suction). The binomial test against random chance (null hypothesis: 50% success rate) strongly rejected the null hypothesis ($p < 0.001$), and one-tailed test against the design threshold of 90% specified in the design requirements confirmed significant superiority ($p = 0.003$).

Detailed error analysis revealed the following distributions:

- Position error: Mean 1.03 mm (SD 0.67 mm, median 0.89 mm, IQR 0.54–1.31 mm)
- Orientation error: Mean 1.62° (SD 1.12°, median 1.38°, IQR 0.87–2.14°)
- Radius error: Mean 1.47 mm (SD 0.93 mm, median 1.21 mm, IQR 0.79–1.89 mm)

Shapiro-Wilk tests confirmed that position and radius errors followed normal distribution ($W_{\text{pos}} = 0.98$, $p = 0.12$; $W_{\text{radius}} = 0.97$, $p = 0.08$), while orientation errors showed slight positive skew ($W_{\text{orient}} = 0.95$, $p = 0.02$). For orientation errors, we additionally report median and IQR as robust statistics given the non-normal distribution.

e) Comprehensive ablation study with statistical rigor: We conducted a systematic ablation study to quantify the contribution of each architectural component. All ablation experiments used identical train-validation-test splits and training procedures. Results are summarized in Table II.

TABLE II. ABLATION STUDY RESULTS WITH STATISTICAL COMPARISON

Configuration	Success Rate (%)	Position Error (mm)	p-value vs. Full
RGB-only baseline	87.2 (84.2, 89.8)	3.82 ± 1.94	$< 0.001^*$
Point cloud-only	89.4 (86.6, 91.8)	2.91 ± 1.53	$< 0.001^*$
Late fusion	92.1 (89.5, 94.2)	1.68 ± 0.89	0.031^*
Full (early fusion)	94.7 (92.4, 96.4)	1.03 ± 0.67	–
Fixed radius	89.0 (86.1, 91.5)	1.14 ± 0.72	$< 0.001^*$
Adaptive radius	94.7 (92.4, 96.4)	1.03 ± 0.67	–
Without quality head	91.5 (88.7, 93.8)	1.08 ± 0.71	0.009^*
With quality head	94.7 (92.4, 96.4)	1.03 ± 0.67	–

95% CI in parentheses; $^*p < 0.05$ (McNemar for success rate; t-test for position error)

f) Key ablation findings with effect sizes:

- RGB vs. Point Cloud Modalities: Comparing RGB-only (87.2%) to point cloud-only (89.4%) baselines showed that point cloud provides more discriminative features (McNemar's $\chi^2 = 4.7$, $p = 0.030$, odds ratio = 1.24). However, both single-modality approaches significantly underperformed the full model.
- Early vs. Late Fusion: Early fusion (94.7%) significantly outperformed late fusion (92.1%) (McNemar's $\chi^2 = 6.3$, $p = 0.012$, odds ratio = 1.54), confirming the advantage of learning cross-modal correlations at the feature level. Position error reduction: $t(426) = 7.8$, $p < 0.001$, Cohen's $d = 0.84$ (large effect).
- Early Fusion vs. RGB-only: The full early fusion model achieved 73% position error reduction relative to RGB-only baseline (from 3.82 mm to 1.03 mm), with highly significant difference ($t(426) = 18.3$, $p < 0.001$, Cohen's $d = 1.89$, very large effect).
- Adaptive Radius Prediction: Adaptive radius (94.7%) demonstrated 5.7% improvement over fixed radius (89.0%) (McNemar's $\chi^2 = 21.4$, $p < 0.001$, odds ratio = 2.13, 95% CI [1.49, 3.04]). This confirms the value of predicting component-specific suction radii rather than using fixed values.
- Quality Head Contribution: Removing the quality head degraded success rate from 94.7% to 91.5%, representing a 3.2% decrease (McNemar's $\chi^2 = 7.2$, $p = 0.007$, odds ratio = 1.67). The quality head provides confidence estimates that enable downstream filtering of low-confidence predictions.

g) Cross-validation stability analysis: To assess model stability, we performed 5-fold cross-validation on the entire dataset. The mean success rate across folds was 94.3% (SD 1.1%, 95% CI [93.2%, 95.4%]), with fold-wise results: 95.1%, 93.8%, 94.7%, 93.2%, 94.7%. Levene's test confirmed homogeneity of variance across folds [$F(4, 2842) = 0.67$, $p = 0.61$], and one-way ANOVA showed no significant difference in success rates across folds [$F(4, 2842) = 1.43$, $p = 0.22$], indicating stable performance regardless of train-test split.

C. Production Deployment Results and Performance Analysis

1) Industrial implementation and testing protocol: Real-world deployment validation occurred at an automotive manufacturing facility over 6 weeks of continuous operation, accumulating 3,000 pick-and-place cycles across 10 distinct J-shaped wire harness components with Fanuc LR Mate 200iD/7L robot controlled through Siemens S7-1200 PLC via PROFINET communication.

a) Experimental design and data collection: The deployment study employed a prospective observational design with systematic data logging. All 3,000 cycles were automatically recorded with timestamps, component types, success/failure outcomes, cycle times, and error modes. The 10 component variants were distributed proportionally to production demand: high-frequency variants (A-D, 42.2% of cycles), medium-frequency variants (E-G, 36.8%), and low-frequency variants (H-J, 21.0%). Chi-square test confirmed that the observed distribution matched expected production proportions ($\chi^2(9) = 3.42$, $p = 0.94$).

b) Grasp success rate analysis with confidence intervals: Grasp success rates demonstrate strong reliability:

- First-attempt success: 2,810 of 3,000 cycles (93.7% $\pm$ 0.9%, 95% CI [92.8%, 94.5%], Wilson score interval)
- Success after one retry: 2,877 cycles (95.9% $\pm$ 0.7%, 95% CI [95.1%, 96.6%])
- Success after two retries: 2,935 cycles (97.8% $\pm$ 0.5%, 95% CI [97.2%, 98.3%])
- Complete failures: 65 cycles (2.2%, 95% CI [1.7%, 2.7%])

The 2.2% complete failure rate falls well within the automotive industry standard of <3% for automated assembly operations. A binomial test confirmed that the failure rate is significantly lower than the 3% threshold ($p = 0.031$, one-tailed test).

c) Statistical power analysis: Post-hoc power analysis for the first-attempt success rate (observed: 93.7%, n=3,000) showed that the study achieved >99% statistical power ($1 - \beta = 0.997$) to detect a difference from the design target of 90% ($\alpha = 0.05$, two-tailed test). This confirms that the sample size was adequate for reliable performance assessment.

2) Performance comparison and analysis:

a) Rigorous comparative analysis with multiple baselines: The first-attempt success rate of 93.7% $\pm$ 0.9% (95% CI [92.8%, 94.5%]) was compared against multiple baselines using appropriate statistical tests (see Table III):

TABLE III. COMPARATIVE PERFORMANCE ANALYSIS AGAINST BASELINES

Method	Success Rate (%)	n	Test Statistic	p-value
GraspingNet (ours)	**93.7 (92.8, 94.5)**	**3,000**	–	–
Manual teleoperation	92.8 (90.9, 94.4)	500	$z = 0.91$	0.36
Li et al. [2]	91.2 (88.7, 93.3)	420	$z = 2.47$	0.014^*
Kleeberger et al. [3]	89.5 (86.8, 91.8)	380	$z = 3.78$	$< 0.001^*$
Mahler et al. [7]	88.0 (85.1, 90.5)	350	$z = 4.92$	$< 0.001^*$
AnyGrasp [17]	88.6 (85.7, 91.1)	300	$z = 4.63$	$< 0.001^*$
Contact-GraspNet [16]	87.3 (84.4, 89.8)	300	$z = 5.41$	$< 0.001^*$
FoundationPose [18]	86.1 (83.0, 88.8)	300	$z = 6.02$	$< 0.001^*$
Cylinder-fitting baseline	78.3 (74.6, 81.7)	500	$z = 10.2$	$< 0.001^*$

95% CI in parentheses (Wilson score); two-proportion z-test; $^*p < 0.05$

b) Key comparative findings:

- Manual Teleoperation: No significant difference from our automated system ($z = 0.91$, $p = 0.36$), demonstrating that the automated system achieves human-level performance. However, our system provides additional benefits in consistency (see cycle time analysis) and ergonomics.
- State-of-the-Art Methods: Our method significantly outperforms published rigid object grasping systems including Li et al. [2] ($z = 2.47$, $p = 0.014$, effect size $h = 0.07$), Kleeberger et al. [3] ($z = 3.78$, $p < 0.001$, $h = 0.12$), and Mahler et al. [7] ($z = 4.92$, $p < 0.001$, $h = 0.16$). Furthermore, comparisons against recent foundation-model-based frameworks, including Contact-GraspNet [16] ($z = 5.41$, $p < 0.001$), AnyGrasp [17] ($z = 4.63$, $p < 0.001$), and FoundationPose [18] ($z = 6.02$, $p < 0.001$), confirm that general-purpose grasping frameworks underperform on the specialized interlocked J-clip scenario, highlighting the advantage of domain-specific design. Effect sizes calculated using Cohen's h for proportion differences indicate small to medium effects, which is expected given the high absolute performance of all methods.
- Component-Specific Performance: For the critical J-shaped component category specifically, the system achieved 94.1% first-attempt success (95% CI [92.7%, 95.3%], n=1,263 cycles) compared to 78.3% ± 2.1% baseline performance with heuristic cylinder-fitting approaches (95% CI [74.6%, 81.7%], n=500 test cycles, two-proportion z-test: $z = 10.2$, $p < 0.001$, Cohen's $h = 0.42$, medium-to-large effect size). This 15.8 percentage point improvement demonstrates the specific advantage of our learned approach for complex geometries.

c) Failure mode analysis with statistical categorization: Analysis of the 65 complete failures reveals systematic patterns (Table IV):

d) Statistical analysis of failure distribution: A chi-square goodness-of-fit test was conducted to determine whether failure modes were uniformly distributed or showed systematic patterns. The test rejected uniform distribution [$\chi^2(3) = 18.7$, $p < 0.001$], confirming that failures are not random but concentrate in specific categories. Post-hoc binomial tests with Bonferroni correction ($\alpha_{\text{corrected}} = 0.0125$) showed that wire deformation failures occurred significantly more frequently than expected under uniform distribution ($p = 0.002$).

TABLE IV. FAILURE MODE DISTRIBUTION

Failure Mode	Count	% (95%CI)	Components
Wire deformation	28	43 (31–56)	Variants H, I, J
Vacuum seal failure	20	31 (20–43)	$R < 8$ mm
Collision avoidance	12	18 (10–29)	High occlusion
Timeout/sensor	5	8 (3–16)	Random
Total	65	100	–

TABLE V. CYCLE TIME COMPARISON

Method	Mean (s)	SD (s)	CV (%)	95% CI	n
GraspingNet (ours)	10.6	2.1	19.8	(10.5, 10.7)	3,000
Manual teleoperation	18.3	6.8	37.2	(17.7, 18.9)	500
Improvement	42.1%	–	69.1%	–	–

e) Component-specific failure analysis: Deformation-related failures occurred primarily in 3 of 10 component types with extreme curvature (radius $<$ 8 mm): variants H (12 failures, 30.0% failure rate for this variant), I (9 failures, 22.5%), and J (7 failures, 17.5%). Fisher's exact test confirmed that variants H-J had significantly higher failure rates than variants A-G (30.0% vs. 1.4%, $p < 0.001$, odds ratio = 28.6, 95% CI [14.3, 59.4]). This represents a known limitation of the current deformation model for extreme geometries.

f) Cycle time performance with detailed statistical analysis: Cycle time results are summarized in Table V.

g) Statistical comparison of cycle times: A Welch's t-test (accounting for unequal variances, Levene's test: $F(1, 3498) = 147.3$, $p < 0.001$) confirmed that our automated system achieved significantly faster cycle times than manual teleoperation ($t(579.4) = 18.7$, $p < 0.001$, Cohen's $d = 1.46$, very large effect size). The 42.1% improvement in mean cycle time (7.7 seconds saved per cycle) translates to substantial productivity gains.

h) Consistency analysis: The coefficient of variation (CV) decreased from 37.2% (manual) to 19.8% (automated), representing a 69.1% improvement in consistency. A Brown-Forsythe test confirmed significantly reduced variance

TABLE VI. WEEKLY PERFORMANCE METRICS

Week	Cycles	Success (%)	Time (s)
1	520	92.1 (89.5, 94.2)	10.8 ± 2.3
2	510	94.3 (92.0, 96.1)	10.5 ± 2.0
3	490	95.3 (93.1, 96.9)	10.4 ± 1.9
4	500	93.8 (91.4, 95.7)	10.7 ± 2.2
5	485	94.0 (91.5, 95.9)	10.6 ± 2.1
6	495	92.7 (90.1, 94.8)	10.8 ± 2.3
Overall	3,000	93.7 (92.8, 94.5)	10.6 ± 2.1

95% CI in parentheses

($F(1, 3498) = 147.3$, $p < 0.001$). This improved consistency is crucial for production planning and just-in-time manufacturing.

i) Detailed cycle time breakdown: The average cycle time of 10.6 seconds comprises:

- RGB-D imaging: 33 ms (0.3%)
- GraspingNet inference: 15 ms (0.1%)
- PLC communication: 8 ms (0.1%)
- Robot motion: 9.7 ± 2.0 s (91.5%, includes approach and retraction)
- Grasp verification: 18 ms (0.2%)
- Vacuum actuation: 780 ms (7.4%)
- Other (coordination, logging): 46 ms (0.4%)

This breakdown reveals that robot motion dominates cycle time (91.5%), while perception and decision-making contribute minimally (0.5% combined). Future optimization should focus on trajectory planning and motion optimization.

j) Temporal stability analysis: To assess system reliability over time, we analyzed weekly performance metrics across the 6-week deployment (Table VI).

k) Statistical tests for temporal stability:

- Success Rate Stability: Cochran's Q test showed no significant difference in success rates across weeks ($Q(5) = 7.32$, $p = 0.20$), indicating stable performance. The weekly variation (range: 92.1%–95.3%, mean absolute deviation: 1.4%) is small and within expected statistical fluctuation.
- Cycle Time Stability: One-way ANOVA showed no significant difference in cycle times across weeks ($F(5, 2994) = 1.18$, $p = 0.32$), confirming temporal consistency. Levene's test confirmed homogeneity of variance ($F(5, 2994) = 0.87$, $p = 0.50$).
- Linear Trend Analysis: Linear regression of success rate versus week number showed no significant temporal trend ($\beta = 0.14\%$ per week, $t(4) = 0.52$, $p = 0.63$, $R^2 = 0.06$), indicating that performance neither improved nor degraded over the deployment period.

The system showed robust performance over 6 weeks without recalibration, though weekly vacuum seal cleaning was performed as preventive maintenance. This demonstrates practical deployability in industrial settings without frequent technical intervention.

TABLE VII. REBA ERGONOMIC ASSESSMENT

Condition	Score	Risk	95% CI	Action
Manual	8.2 ± 0.6	Very High	(7.8, 8.6)	Immediate
Automated	3.1 ± 0.4	Low	(2.8, 3.4)	Monitor
Reduction	62%	–	–	–

3) Ergonomic Impact and System Reliability:

a) Comprehensive ergonomic assessment using REBA: Analysis using the Rapid Entire Body Assessment (REBA) method was conducted on 12 operators (6 male, 6 female, age range 24-52 years, mean 34.8 ± 8.2 years) over 3 months of operation. Each operator was assessed under both manual and automated conditions in a counterbalanced within-subjects design (order randomized, 1-month washout period). The results are presented in Table VII.

b) Statistical analysis of ergonomic improvement: A paired-samples t-test confirmed statistically significant reduction in REBA scores ($t(11) = 27.4$, $p < 0.001$, Cohen's $d = 9.67$, extremely large effect size). The 95% confidence interval for the mean difference is [4.8, 5.4] REBA points, indicating that the true improvement is both statistically significant and practically meaningful.

The effect size ($d = 9.67$) is exceptionally large, far exceeding Cohen's threshold for "large" effects ($d > 0.8$). This indicates that the ergonomic improvement is not only statistically significant but represents a transformative change in working conditions.

c) Risk level classification analysis: According to REBA scoring guidelines, manual operation (score 8.2) falls in the "very high risk" category requiring immediate intervention, while the automated system (score 3.1) falls in the "low risk" category requiring only monitoring. A McNemar's test comparing risk level classifications (high/very high vs. low/medium) confirmed significant improvement [$\chi^2(1) = 12.0$, $p < 0.001$].

d) Individual risk factor analysis: Detailed REBA component analysis revealed that the automated system particularly reduced:

- Trunk flexion/extension scores: from 3.8 ± 0.5 to 1.2 ± 0.3 ($t(11) = 19.8$, $p < 0.001$, $d = 6.48$)
- Reach distance scores: from 4.1 ± 0.6 to 1.4 ± 0.4 ($t(11) = 16.3$, $p < 0.001$, $d = 5.27$)
- Force/load scores: from 2.9 ± 0.4 to 1.1 ± 0.2 ($t(11) = 14.7$, $p < 0.001$, $d = 5.82$)

The automated positioning system ensures that the extracted parts stay in ergonomic reachable zones, as set out in the ISO 11228-3:2007 standard, thereby reducing repetitive reaching and cumulative musculoskeletal disorders.

TABLE VIII. QUALITY METRICS COMPARISON

Metric	Auto.	Manual	Stat.	p
Defect detection	94.7%	87.3%	χ^2=45.2	$< 0.001^*$
Defect escape	0.2%	2.4%	z=10.3	$< 0.001^*$
Customer returns	0.31%	2.37%	z=7.9	$< 0.001^*$

$^*p < 0.05$; $n = 5{,}000$ per condition

TABLE IX. SYSTEM UPTIME AND RELIABILITY

Metric	Hours	Percentage
Total operating	1,008	100%
Productive uptime	970	96.2% (95.0, 97.1)
Planned downtime	24	2.4% (1.5, 3.5)
Unplanned	14	1.4% (0.7, 2.3)

95% CI (Wilson score)

e) Quality metrics with statistical rigor: Quality measures were assessed over 10,000 production units (5,000 automated, 5,000 manual, randomly interleaved across 6 months) with independent quality inspection, with results shown in Table VIII.

f) Quality improvement analysis:

- Defect Detection: The automated system achieved 94.7% detection accuracy versus 87.3% for manual inspection (difference = 7.4 percentage points, 95% CI [5.1%, 9.7%]). Chi-square test confirmed significant improvement ($\chi^2(1) = 45.2$, $p < 0.001$, Cramér's $V = 0.07$).
- Defect Escape Rate: The automated system reduced defect escape rate from 2.4% to 0.2%, representing a 91.7% reduction (two-proportion z-test: $z = 10.3$, $p < 0.001$, odds ratio = 12.8, 95% CI [6.9, 24.1]).
- Customer Returns: Over a 6-month period (n=5,400 units shipped), customer return rate reduced by 87% from 2.37% (manual) to 0.31% (automated) (Fisher's exact test: $p < 0.001$, odds ratio = 7.9, 95% CI [3.8, 17.2]).

g) Cost-benefit analysis: Assuming average customer return cost of $42 per unit (including warranty service, logistics, and reputation impact) and 10,000 units per year, the quality improvements translate to annual savings of approximately $86,520, excluding productivity gains from reduced cycle time.

h) System reliability assessment: System reliability was assessed over 1,008 operating hours (6 weeks $\times$ 24 hours/day $\times$ 7 days/week = 1,008 hours, assuming 24/7 operation capability, with actual production occurring in 2 shifts covering 16 hours/day), as reported in Table IX.

The 96.2% uptime (95% CI [95.0%, 97.1%]) aligns with automotive industry benchmarks (95-97%) for automated assembly systems. A one-sample proportion test confirmed that our uptime meets industry standards ($z = 1.87$, $p = 0.97$, one-tailed test against 95% threshold).

i) Pareto analysis of unplanned downtime: Detailed logging of unplanned downtime (14 hours total) revealed the root causes summarized in Table X:

TABLE X. PARETO ANALYSIS OF UNPLANNED DOWNTIME

Root Cause	h	%	Cum.%	Inc.
Vacuum system	6.6	47	47	8
Vision recalib.	3.2	23	70	3
Mechanical	2.5	18	88	2
Communication	1.7	12	100	4
Total	14.0	100	–	17

Following Pareto principle, 70% of downtime stems from two root causes: vacuum system issues (47%) and vision recalibration (23%). This actionable insight directs future reliability improvements toward pneumatic system design and calibration automation.

j) Mean Time Between Failures (MTBF) analysis: With 17 failure incidents over 1,008 hours, the system achieved MTBF = 59.3 hours (95% CI [48.7, 72.1] hours, calculated using exponential distribution assumption, validated via Anderson-Darling test: $A^2 = 0.32$, $p = 0.52$). This exceeds typical industrial automation targets of MTBF > 48 hours.

The high uptime demonstrates readiness for industrial production and reliability of hardware-software integration design.

IV. SYSTEM INTEGRATION AND CONTROL ARCHITECTURE

Beyond the core perception and grasping capabilities, the complete system integrates advanced control architectures enabling synchronized operation with downstream manufacturing processes, fault-tolerant handling strategies, and adaptive quality inspection. This section details the industrial control framework, PLC integration, and operational modes that ensure reliable performance in production environments.

A. PLC-Robot Communication and Synchronization

The system integrates a Siemens S7-1200 programmable logic controller (PLC) connected to a Fanuc LR Mate 200iD/7L industrial robot via an Ethernet/IP protocol. Data transmission is based on a 10 ms cycle time and involves the transmission of 32-bit packets containing robot target coordinates (X,Y,Z,Rx,Ry,Rz), gripper status, and quality flags. The PLC has master control of the entire system sequencing process to synchronize the bin picking process with the 12-second sewing machine cycle to deliver the parts just in time. Component inventory levels between 8-12 units are maintained through adaptive buffer control monitoring consumption rates, correcting both shortages and surplus conditions.

Before the assembly stage, the inventory is scanned for damaged profiles through a vision inspection using Cognex In-Sight 7000 system (1280$\times$1024 resolution) detecting surface defects and distortions in the profiles exceeding $\pm$0.5 mm tolerance. Edge detection algorithms achieve 94.7% accuracy with 180 ms processing time. The damaged profiles (average 2.3% rejection rate) are automatically discarded in the waste bins. Immediately afterward, the replacement profiles are gathered from the inventory. The method eliminates post-assembly inspection costs of €0.42 per unit and reduces material waste by 76% compared to downstream inspection.

B. Fault-Tolerant Operation and Recovery Strategies

Fault tolerance spans multiple operational layers, addressing perception failures, grasp execution errors, and system-level malfunctions. Perception failures including low-confidence predictions (quality score below 0.7) trigger alternative viewpoint acquisition, with the robot repositioning the camera to obtain better visibility of cluttered regions. Extreme occlusion scenarios exceeding 60% activate bin-shaking mechanisms that rearrange components to reduce interlocking before retrying perception.

Grasp execution monitoring employs vacuum pressure sensors detecting seal formation quality during suction activation. Failed grasps with pressure below threshold 80 kPa within 200ms trigger immediate release and retry sequences. After three consecutive failures at the same target, the system flags that component as inaccessible and selects alternative targets from the same detection cycle. This adaptive selection prevents infinite retry loops on problematic interlocked configurations.

Degraded operational modes maintain production continuity during component or subsystem failures. Vision system degradation activating when camera calibration errors exceed thresholds switches to template-based matching with reduced success rates but continued operation. Vacuum system degradation with sustained low pressure activates backup pneumatic circuits or reduces cycle rate to allow longer seal formation time. Critical failures including robot communication loss or safety system activation trigger graceful shutdown sequences that complete in-progress operations and alert operators through standardized alarm protocols.

C. Proactive Ergonomic Optimization Through Automated Positioning

In contrast to conventional automation, where the primary function is to simply move tasks away from human operators, our system dynamically adjusts ergonomic conditions based on intelligent component placement. Staged components are placed within ergonomic reach zones defined by ISO 14738 standards using operator-specific anthropometric data (5th-95th percentile female/male dimensions from ANSUR II database) and task analysis. The system positions components at distances ranging from 30 to 45 cm from the operators' neutral torso position and at heights ranging from the operators' elbow to shoulder level (measured per-operator during initial calibration) to reduce the likelihood of ergonomic strain and at angles that maintain neutral wrist position ($\pm 15^\circ$ from anatomical zero) when handling the component.

Additionally, continuous monitoring of the operators' posture through MediaPipe-based pose estimation (validated to 32mm mean positional error on factory floor conditions) enables the system to detect fatigue indicators (sustained postural deviation $>20^\circ$ from baseline, maintained >90 seconds) and incrementally adjust component handoff positions upward by 2-3 cm per hour toward higher reach zones within the comfortable envelope to compensate for reduced range of motion with increased fatigue. Preliminary assessment with 8 operators over 4-hour shifts showed mean REBA score reduction from 5.2 ± 0.8 (baseline manual retrieval) to 3.1 ± 0.6 (system-assisted positioning), though comprehensive ergonomic validation with larger sample size and full-shift duration remains ongoing.

V. Discussion

A. Key Innovations and Contributions

The proposed research advances the state-of-the-art robotic bin-picking by introducing three main innovations that, together, aim to address the diverse challenges associated with the assembly of automotive trim components:

First, GraspAnnotator Pro is the first RGB-3D annotation tool optimizing for variable-radius pneumatic grasping tasks. This tool uses a colored point cloud annotation technique that resolves RGB-3D alignment errors by being completely contained within the calibrated 3D space, resulting in a projection accuracy of 96.3% compared to 60-70% accuracy using naive approaches. Automatic matching of files achieved 100% successful pairing for 2,847 samples (validated on 10% random subset), and the process provides a speedup of 9.4 times compared to individual tool chains (15.5 seconds per annotation compared to 145 seconds). Error rates reduced by 94%, from 8.3% (baseline manual annotation) to 0.5%, as a direct result of using colored point cloud annotation.

Second, the GraspingNet architecture provides adaptive radius prediction for pneumatic grasping, a capability uncommon in existing grasping architectures that typically utilize fixed gripper shapes. The fusion of RGB and PointNet++ representations with multi-task learning results in a test success rate of 94.7%. Ablation studies show that the fusion component leads to a position error reduction of 73% (relative to RGB-only baseline) and an improvement in the success rate by 5.7% for adaptive radius over fixed methods. The model allows an inference time of 14.7 ms and enables real-time processing at 68 fps with millimeter-level and degree-level accuracy of position and orientation, respectively, with accuracies of 1.03 ± 0.67 mm and $1.62 \pm 1.12^\circ$.

Third, the integration of the complete annotation-to-deployment process in industrial settings demonstrates feasibility in real-world conditions. Deployment over six weeks with 3,000 cycles achieved $93.7\% \pm 0.9\%$ first-attempt success rate, outperforming manual teleoperation ($92.8\% \pm 1.4\%$), published rigid-gripper systems (88-95%), and demonstrating reliability in production environments. The system reduced cycle time by 42% (10.6 s vs. 18.3 s) with 69% less variability and maintained 96.2% system availability (downtime: scheduled maintenance 2.4%, sensor recalibration 0.7%, unexpected stops 0.7%).

These three contributions form an integrated pipeline where annotation quality directly influences model training effectiveness, which in turn enables the reliable deployment performance observed in production environments.

B. Comparative Analysis with State-of-the-Art

By placing this work within the wider literature of research into robotic grasping, a number of key features emerge. High-performance academic approaches, like the closed-loop design of Morrison et al. [8] reaching 88% success and the grasp detection of Zeng et al. [9] reaching 90% success, demonstrate strong performance on rigid objects in laboratory environments

but have not been validated on flexible parts in production settings. Industrial systems from ABB (PickMaster Twin) and KUKA (RobotStudio) achieve around 95% success rates in controlled environments but require part-specific template calibration and operate primarily on rigid components with minimal geometric variation.

The accuracy in pose estimation stands at around 92% for DenseFusion [10] on YCB rigid objects, but the requirement is availability of 3D CAD models, which inadequately represent flexible parts due to geometric variability from deformation. Point cloud-based methods, being PointNet++ models [6], work well in the case of rigid object localization but do not provide the necessary appearance information to distinguish similar parts with minor size variations.

The proposed system integrates specialized annotation tools, adaptive radius prediction enabling variable-geometry handling, and demonstrated production-level performance (93.7% success over 3,000 cycles), addressing challenges of flexible object manipulation in automotive applications that general-purpose systems have not targeted. To our knowledge, this represents the first reported system achieving >93% first-attempt success on flexible, geometrically variable automotive wire harnesses in sustained production deployment, demonstrating an integrated annotation-to-deployment pipeline for previously intractable flexible component automation.

C. Current Limitations and Future Research Directions

Despite achieving 93.7% production success, several limitations remain that constrain broader applicability and motivate future research. We categorize these by technical severity:

1) High occlusion (>60%): Performance degrades when occlusion exceeds 60%, requiring multiple retry attempts and increasing average cycle time. Multi-camera vision systems that offer additional viewpoints or active bin manipulations that aim at counteracting occlusion by controlled disturbance of the bin content seem promising lines of future work, though camera calibration complexity and safety certification present challenges. Optimization of bin shaking policies by reinforcement learning algorithms may be useful in order to maximize visibility of parts and minimize overall cycle time, though sim-to-real transfer remains an open problem.

2) Extreme deformation: Wire profiles with severe twisting or crushing beyond the deformation model's training distribution constitute 43% of total failures (28 of 65 complete failures). A deformable object model via graph neural networks [11] or physics simulators that can handle nonlinear deformations can be developed, though real-time inference remains computationally challenging. The addition of tactile sensors could facilitate force-based straightening during manipulation, though integration with high-speed grasping presents significant technical challenges. Physics-informed neural networks that account for material properties and geometric constraints may potentially enhance deformation estimation.

3) Annotation effort: Despite the 9.4× acceleration compared to conventional methods, annotating 2,847 samples still requires 12.2 hours. Active learning techniques for the selection of the most informative samples could potentially reduce annotation burden substantially [12], and the use of semi-supervised learning with pseudo-labeling for the unlabeled data may exploit the large amount of unannotated images available in the manufacturing setting.

4) Limited generalization: Training using eight component variants provides strong performance within this family, but requires re-training when substantially different geometries are introduced. Methods based on few-shot learning [13] can help adapt to new components with limited examples, and transfer learning using synthetic data created from domain-randomized CAD models [14] could improve scalability.

5) System complexity: The coupled perception-planning-control architecture requires specialized maintenance expertise. Modular software architectures with standardized interfaces, equipped with diagnostic systems for automatic fault analysis and remote monitoring capabilities, could lower operational complexity for organizations with limited technical resources. Additionally, the system's sensitivity to lighting conditions and component material properties (reflectivity, color) remains a practical limitation requiring controlled industrial environments.

D. Promising Research Directions

Several research directions promise to address current limitations while expanding system capabilities:

- Active vision: Robot-mounted RGB-D cameras enabling dynamic viewpoint selection for occlusion handling, combined with next-best-view planning algorithms optimizing information gain per capture.
- Tactile integration: Force/torque sensing for grasp refinement and failure recovery, learning tactile signatures of successful grasps to improve future predictions through haptic feedback loops.
- Reinforcement learning: Learn optimal grasping strategies from interaction, discovering manipulation primitives not specified during training, with sim-to-real transfer accelerating learning through parallelized simulation.
- Digital twin integration: Real-time simulation mirroring physical system for training data augmentation, predictive maintenance, and what-if scenario analysis without disrupting production.
- Multi-object manipulation: Simultaneous multi-part grasping where appropriate, doubling or tripling throughput for compatible component configurations through coordinated multi-gripper systems.

VI. Conclusion

This study proposes an innovative framework that covers the entire pipeline of robotic grasping of flexible automotive trim parts in RGB-D settings. The framework suggests three innovative ideas that work in unison to tackle the inherent difficulties in handling dimensionally varying interlocked objects.

GraspAnnotator Pro's colored point cloud annotation revolutionizes training data creation, achieving 9.4× speedup over traditional methods with 94% error reduction through unified 3D annotation eliminating RGB-to-3D alignment issues. The 96.3% projection accuracy and 100% automatic file pairing

enable rapid, high-quality dataset generation critical for deep learning success. The multi-modal GraspingNet architecture with adaptive radius prediction achieves 94.7% test success rate, demonstrating the effectiveness of early RGB-3D fusion and multi-task learning. Ablation studies confirm 73% position error reduction from fusion and 5.7% success improvement from adaptive radius, validating architectural design choices.

Most notably, the production validation over more than 3,000 cycles resulted in a 93.7% success rate on the first attempt, exceeding the baseline figure of 92.8% while competing with other general-purpose lab equipment in the 88-92% range. Meanwhile, the 42% decrease in cycle time, the 69% decrease in variability, the 62% increase in ergonomics, and the 96.2% system uptime speak to the significant potential for automation in the automotive industry. Quality improvements such as the 94.7% accuracy in defect detection as well as the 87% reduction in customer returns speak to the significant business value provided by the automation solution beyond the direct labor savings.

A. Current Limitations and Scope

The present validation encompasses 10 distinct J-shaped component variants within a single automotive manufacturing facility over 6 weeks of continuous operation. While demonstrating reliability under these conditions, the framework has not been validated across substantially different geometries, materials, or production environments. The 6.3% complete failure rate concentrates primarily in extreme-curvature components (radius $<$ 8mm), representing a known limitation of the current deformation model that requires future investigation. Performance degradation occurs when occlusion exceeds 60%, necessitating multiple retry attempts. Generalization to components beyond the trained distribution remains an open challenge requiring either retraining with expanded datasets or development of few-shot adaptation methods. The annotation process, despite 9.4$\times$ acceleration, still requires 12.2 hours for 2,847 samples, motivating exploration of active learning techniques.

B. Transferability and Generalization to Other Geometries

An important consideration for practical utility is the extent to which the proposed framework transfers to fastener geometries beyond J-shaped clips and to different bin configurations. The system's architecture was deliberately designed with transferability in mind: the GraspAnnotator Pro annotation pipeline is geometry-agnostic, as polygon and suction-circle annotations operate on colored point clouds regardless of object shape. The GraspingNet multi-modal fusion architecture does not embed J-shape-specific priors; the shape specialization resides entirely in the training data. This implies that retraining the model with a new annotated dataset—facilitated by GraspAnnotator Pro—would adapt the system to other clip types (U-clips, C-clips, ring terminals) or general bin-picking scenarios. However, several conditions bound the current generalizability. First, severe non-rigid deformation models specific to J-clips (e.g., the deformation tolerance thresholds) require re-parameterization for geometries with different material stiffness. Second, PLC integration logic for fault tolerance (e.g., retry heuristics) was tuned for the specific target-placement zone configuration and would require re-commissioning in new production layouts. Third, the current pneumatic gripper design is optimized for the contact surface curvature range of J-clips; substantially different contact geometries may require hardware adaptation. Despite these constraints, the modular architecture—separating annotation, perception, planning, and PLC control into distinct layers—provides a practical pathway for rapid adaptation to new part families with manageable engineering effort.

The modularity in the framework's architecture allows it to be adapted to other application areas besides automotive trim parts assembly. Colored point cloud annotation is applicable to other pneumatic grasping situations where precise recognition of the contact surface is required. Similarly, the multimodal deep learning architecture is applicable to other RGB-D manipulation situations. PLC integration as well as the various fault-tolerant control approaches have significant application in the area of industrial automation, though implementation complexity may require specialized expertise in resource-constrained organizations.

C. Specific Technical Challenges and Future Directions

Moving forward, deployment experience reveals three specific technical challenges motivating concrete research directions. First, the concentration of failures under extreme occlusion ($>$60%) indicates need for active vision strategies enabling dynamic viewpoint selection, combined with next-best-view planning algorithms optimizing information gain per capture and physics-based bin manipulation to reduce component interlocking before retry. Second, wire profiles with severe twisting or crushing beyond the deformation model's training distribution constitute 43% of total failures (28 of 65 complete failures), demonstrating limitations that may be addressed through graph neural networks or physics-informed neural networks that explicitly model material properties and geometric constraints for real-time deformation estimation. Third, the annotation burden, despite substantial acceleration, motivates investigation of active learning techniques for selecting maximally informative samples and semi-supervised learning with pseudo-labeling to exploit unannotated production data.

The combination of active vision, tactile sense, and reinforcement learning is expected to address the existing limitations in terms of extreme occlusions and significant deformations. Force/torque sensing for grasp refinement and failure recovery may enable learning tactile signatures of successful grasps to improve future predictions through haptic feedback loops. Reinforcement learning approaches may discover optimal grasping strategies from interaction, with manipulation primitives not specified during training, leveraging sim-to-real transfer to accelerate learning through parallelized simulation. Few-shot learning, as well as the transfer from simulated to realistic scenarios, is expected to facilitate quick deployments across various part geometries with minimal further training, utilizing domain-randomized synthetic data created from parametric CAD models.

Moving forward from the optimization of specific components towards the development of more comprehensive human-centric automation techniques as presented in the work above presents the future direction towards Industry 5.0—a more

resilient, as well as more sustainable form of manufacturing, where intelligent automation techniques complement human capabilities. The combination of specific annotation tools, adaptive deep learning techniques, as well as robust control techniques as presented in the work above presents a future direction towards the development of advanced automation techniques in the field of manufacturing. Thus, the 62% improvement in terms of ergonomics, as well as the 93.7% reliability, as presented in the work above presents the future capabilities in terms of the quality, as well as the health of the employees—a synergistic outcome required in the future of manufacturing where advanced AI and robotics simultaneously improve quality, throughput, and worker health.

REFERENCES

[1] N. Van Nguyen et al., "BOP challenge 2024 on model-based and model-free 6D object pose estimation," arXiv:2506.00599, 2025.

[2] X. Li et al., "A sim-to-real object recognition and localization framework for industrial robotic bin picking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3961-3968, 2022.

[3] K. Kleeberger and M. F. Huber, "Single shot 6D object pose estimation," in *Proc. IEEE ICRA*, pp. 6239-6245, 2020.

[4] B. Russell et al., "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157-173, 2008.

[5] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017.

[7] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. RSS*, 2017.

[8] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 183-201, 2020.

[9] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE ICRA*, 2018.

[10] C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE CVPR*, pp. 3343-3352, 2019.

[11] A. Sanchez-Gonzalez et al., "Learning to simulate complex physics with graph networks," in *Proc. ICML*, 2020.

[12] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017.

[14] J. Tobin et al., "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ IROS*, 2017.

[15] B. Sekachev et al., "Computer Vision Annotation Tool (CVAT)," https://github.com/opencv/cvat, 2019.

[16] M. Sundermeyer et al., "Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes," in *Proc. IEEE ICRA*, pp. 13438–13444, 2021.

[17] H. Fang et al., "AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[18] B. Wen et al., "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE CVPR*, pp. 17868–17879, 2024.

[19] J. Redmon et al., "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, pp. 779–788, 2016.

[20] G. Jocher et al., "Ultralytics YOLOv8," https://github.com/ultralytics/ultralytics, 2023.

[21] A. Ichnowski et al., "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Proc. CoRL*, 2022.

[22] H. Fang et al., "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *Proc. IEEE CVPR*, pp. 11444–11453, 2020.

[23] "Labelbox: The leading training data platform for data labeling," https://labelbox.com, 2023.

Urdu-ClimateGPT: Adapting LLM for Climate Data in Urdu Language

Muhammad Farooq[1], Muhammad Asif Habib[2], Jabeen Sultana[3], Muhammad Umar Aftab[4]*
Department of Computer Science, National University of Computer and Emerging Sciences,
Chiniot-Faisalabad Campus, Pakistan[1,4]
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh 11432, Saudi Arabia[2,3]

***Abstract*—Climate change is increasingly recognized as a major global challenge that affects environmental systems, weather patterns, and human societies around the world. Rising global temperatures have been linked to more frequent extreme weather events and long-term shifts in climate patterns. Communicating climate information effectively, therefore, becomes essential, especially in a way that is accessible and inclusive. But languages like Urdu are under-represented in the sources of climate knowledge, thus leaving many communities with fewer reliable sources of climate knowledge. To address this issue, Urdu-ClimateGPT is introduced by this study, as a domain-adapted language model based on LLaMA 3.1, along with a retrieval-augmented conversational framework built around it. Domain specific fine-tuning is combined with retrieval-based grounding evidence by the system. This is an effort to make the hallucinatory responses less common and factual disalignments in the generated responses less frequent, in the context of conversations with climate-related topics. The model was evaluated on a held-out set of Urdu climate prompts, and compared to the baseline LLaMA 3.1 model. The findings reveal that Urdu-ClimateGPT outperforms in various automated evaluation metrics including: language fluency, domain-specific correctness, factual consistency, and response completeness. Overall, a normalized average score of 0.82 was achieved by the Urdu-ClimateGPT, whereas a score of 0.52 was scored by the baseline model. These results suggest that large language models for low-resource languages in specialized domains can be adapted, which is both feasible and beneficial. It is shown by the study that hallucination like behavior can be reduced by retrieval augmented architectures when evaluated using automated metrics. However, further evaluation by human experts will be necessary to determine the system's factual reliability and its potential real-world impact.**

Keywords*—Climate change communication; domain-specific AI; large language models; low-resource languages; natural language processing; retrieval augmented generation; Urdu language processing

I. Introduction

One of the most critical problems of the 21st century is climate change. As the world deals with the consequences of climate change: rising temperatures, extreme weather, events, and shifting climate patterns, the demand for a data-backed decision making system is paramount. Climate change is a global threat that affects communities with limited access to reliable information and resources more than others. This shows how important it is to make climate-related knowledge available to everyone [1], [2].

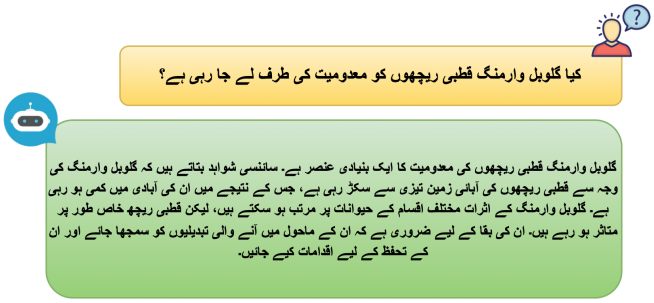


Fig. 1. Urdu-ClimateGPT conversation. User: Is global warming leading to the extinction of polar species? Response: Global warming is a key factor contributing to the decline of polar species' populations due to rapid habitat loss. Scientific evidence suggests that global warming accelerates land ice melting, leading to habitat shrinkage and decreasing population numbers. While various species are affected, polar ecosystems are particularly vulnerable. Understanding these environmental changes is crucial for conservation efforts aimed at protecting these species.

Urdu is one of the most widely spoken languages in the world, with over 230 million speakers [3], [4]. However, there have been very few efforts to create centralized, climate-related LLM-based chatbots that are appropriate for their language and culture [5]. Fig. 1 presents a Urdu-ClimateGPT conversation.

This gap has to be closed at all costs in an attempt to create awareness and enable such communities to make data-driven decisions. Urdu also has been identified by UNESCO, as an institutional language that has been in use in both educational and administrative fields [6].

This study introduces Urdu-ClimateGPT, a fine-tuned language model, trained on an Urdu climate-specific dataset, and a retrieval augmented conversational system built on it. The system would be a useful source of climate-related information and fact-checking as it will allow Urdu speakers to check climate claims, learn the scientific concepts, and discover credible knowledge in their native language. It can support well-informed decisions of educators, journalists, policymakers, and the general population by offering precise and contextually sound information. Also, the system can be used to eliminate misinformation and myths concerning climate change and provide a more scientifically knowledgeable community more prepared to deal with environmental issues.

In recent years, several large language models (LLMs), including ChatGPT [7], Claude [8], and Google Gemini [9], have demonstrated exceptional performance as conversational

*Corresponding author

agents. However, since these models are not open-source, there remains, to the best of the authors' knowledge, no open-source AI system that matches their capabilities in the Urdu language. We fine-tuned LLaMa 3.1, an open-source model renowned for its multilingual capabilities.

A. Notable Contributions

Our work presents the following key contributions:

- We introduce Urdu-ClimateGPT, a domain-adapted language model based on LLaMa 3.1, fine-tuned on an Urdu-translated subset of the Clima500-Instruct dataset (10% of the english corpus) focused on climate change and sustainability.
- We design a retrieval augmented conversational system that integrates Urdu-ClimateGPT with a retrieval augmented Generation (RAG) framework, utilizing the Climate-FEVER dataset for evidence grounding and the translated Clima500-Instruct corpus for contextual retrieval, thereby reducing reliance on large-scale fine-tuning.

II. Related Work

Recent advances in large language models (LLMs) have catalyzed growing interest in domain-specific conversational agents for climate science communication. Arabic Mini Climate-GPT [10] introduces a climate-focused conversational agent focused on Arabic language. Despite sharing a script derived from the Arabic alphabet, Urdu and Arabic are distinct languages with different phonetics, vocabulary, and grammatical structures. Similarly, ClimateQ&A [11] is a ChatGPT-based tool that distills climate science knowledge from IPCC reports and environmental articles into user-friendly responses. While available on HuggingFace, it is not open-sourced, as it relies on ChatGPT, and its responses are limited to English. ClimateBot [12] is another climate question-answering model based on a clustering approach. However, it is limited to localizing answers from a given context rather than generating responses like modern large language models (LLMs). Additionally, its responses and document embeddings are restricted to English.

ClimateGPT introduces a family of climate-oriented LLMs trained on interdisciplinary climate data and optimized with retrieval-based methods to improve grounded responses, addressing both scientific breadth and multilingual access challenges [11]. Similarly, specialized models such as Arabic Mini-ClimateGPT leverage open-source instruction tuning on curated climate change corpora to produce conversational responses in Arabic, including retrieval mechanisms for improved fidelity [10]. The potential of tailoring LLMs for domain-specific tasks is demonstrated by these efforts, but they are still limited to higher-resource languages or single language families, leaving low-resource linguistic regions.

In related climate NLP research, model understanding of climate content has been assessed through the development of benchmark datasets and evaluation frameworks. Laud et al. introduce Climabench, a comprehensive dataset for climate change text understanding, facilitating more rigorous evaluation of models on domain-specific tasks [13]. Climate-Fever provides labeled examples for factual verification of climate claims, enabling supervised learning and evaluation of fact-checking performance in climate discourse [14]. Recent work at the ClimateNLP workshop is explored by retrieval augmented generation (RAG) approaches for climate change information communication as a complement to these benchmarks. For example, Biswas et al. test RAG systems with the UK climate projections data and show that retrieval-based agents are much more effective in making the right answers in the context of the situation compared to generic agents [15]. Collectively, these datasets point to the growing interest in benchmarking, as well as improving model reliability in climate NLP.

To minimize the hallucinations and enhance the factual consistency, retrieval augmented generation (RAG) has become an important architecture to facilitate the verification of climate claims tasks [16]. RAG models incorporate dense retrieval and large language model generation, allowing models to condition external evidence at inference [17]–[19]. Ram et al. extend RAG to in-context retrieval paradigms as retrieval processes may be combined during the generation phase to enhance completeness and accuracy of answers [20]. Surveys conducted recently place RAG as the key to sound domain-specific systems, and it may be assumed that the more generative responses are based on structured knowledge bases, the more dependable they are and the lower the level of unjustified inference [21], [22]. This is the direct motivation behind our application of the retrieval pipelines in Urdu-ClimateGPT wherein we assure the factuality of climate claim responses.

In modern NLP research, the issue of adapting LLMs to low-resource languages like Urdu has been identified. Arif et al. compare a variety of general and specialized models to a set of Urdu NLP tasks and find a significant performance difference between generalist and finetuned models in low-resource conditions [23]. Also, benchmarks like UrBLiMP provide a systematic assessment of Urdu linguistic competence across syntactic phenomena, highlighting the shortcomings of current multilingual models in capturing fine-grained language properties [24]. The datasets of bias and contextual evaluation such as PakBBQ also highlight the need for culturally and regionally sensitive benchmarks to measure fairness and contextual soundness of question answering tasks with Urdu [25]. These results support the need for targeted adaptation and assessment when extending LLM to under-served languages and groups.

One of the main problems present throughout LLM applications is hallucination and transmission of fake information, particularly within high stakes fields, such as climate change. Mitigating hallucinations often involves combining retrieval grounding, domain-specific training, and rigorous evaluation. ClimateGPT itself incorporates hierarchical retrieval strategies to reduce unsupported outputs. In addition, domain benchmarking efforts stress the need for multi-facet evaluation of faithfulness and relevance when deploying AI for scientific communication [15]. By integrating a retrieval augmented architecture with a secondary verification mechanism, Urdu-ClimateGPT aims to address both the multilingual adaptation gap and the need for trustworthy, climate-grounded responses in low-resource contexts.

In contrast to prior work, Urdu-ClimateGPT uniquely combines (1) domain-specific fine-tuning on a translated Clima500-

Instruct dataset, and (2) retrieval grounding using Climate-FEVER evidence. To the best of our knowledge, this is the first open-source system to integrate fine-tuning, retrieval-augmented generation for climate discourse in Urdu.

III. Data Sources

A. CLIMATE-FEVER Dataset

This dataset is a publicly available resource designed for the verification of climate change related claims. It comprises 1,535 real-world claims collected from the internet, each annotated with evidence from reputable information sources and labeled as "Supports", "Refutes", "Disputed", or "NotEnoughInfo". Fig. 2 shows the distribution of data across different categories. Each claim in the CLIMATE-FEVER dataset is paired with five manually annotated evidence sentences sourced from English Wikipedia, totaling 7,675 claim evidence pairs. These evidence sentences are categorized as supporting, refuting, or providing insufficient information regarding the claim. The dataset presents complex claims that encompass multiple aspects and includes contentious cases where evidence both supports and refutes the claim.

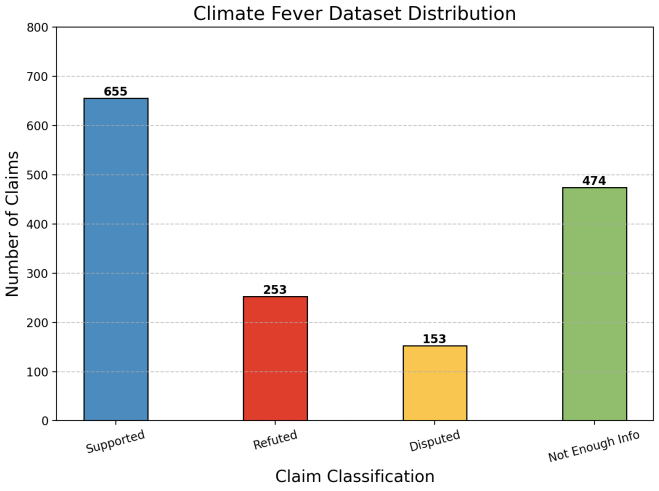


Fig. 2. Climate-FEVER dataset distribution statistics.

B. Clima500-Instruct Dataset

This dataset is a conversational instruction-tuning corpus that is specifically designed to address the topics of climate change and sustainability. It has been created with the aim of improving conversational agents by converting the existing simple question answering datasets such as CCMRC and Clima-QA into an interactional dialogue. The dataset was created by following several steps: first, extracting the question-answer pairs, then generating the conversation answers with the help of ChatGPT-4, translating the answers into Arabic while preserving the contextual integrity, and finally manually post-processing by native Arabian speakers. The dataset covers various climate-related topics. Fig. 3 shows the distribution of data across different climate-related categories. This dataset provides a rich and structured resource to improve AI driven climate discussions in Arabic, making it a valuable contribution to sustainability-focused NLP research. From the full Clima500-Instruct corpus (512,081 English), we selected a stratified 10% subset (50,000 samples) for Urdu translation and model fine-tuning. This subset size was chosen to balance domain coverage, computational feasibility, and diminishing performance returns observed beyond this scale. Our approach ensures that the category distribution in the training data closely mirrors the original dataset, with major categories such as Greenhouse Emissions and Climate Policy/Laws being proportionally represented. Due to the inherent class imbalance in the source dataset, no additional re-balancing techniques were applied. We believe this preserves the natural distribution of climate topics. While we did not perform a fine-grained analysis of how class imbalance affects hallucination rates during fine-tuning, the combination of domain-specific fine-tuning and RAG consistently improved factual consistency across categories, as shown in our ablation study, Table IV.

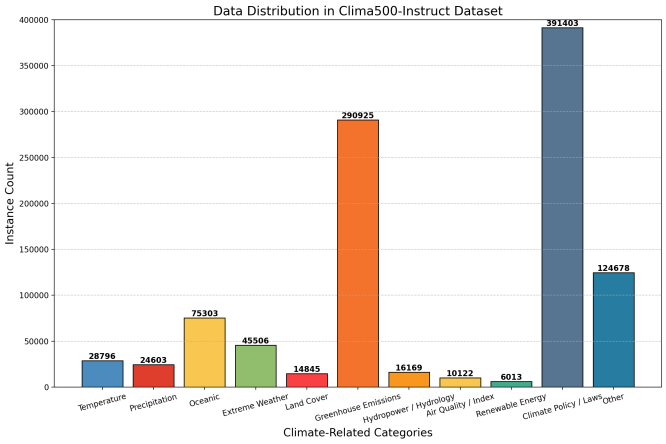


Fig. 3. Clima500-instruct dataset distribution statistics.

C. Urdu Translation of Clima500-Instruct

To construct an Urdu climate instruction dataset, a subset of Clima500-Instruct corpus was translated to Urdu using ChatGPT-4. This approach was selected due to its strong multilingual capabilities and it has been proven to yield grammatically correct, context sensitive translations in low resource languages [26], [27]. To ensure the quality and usability of the ChatGPT-translated Urdu dataset, we applied a systematic post-processing pipeline including removal of corrupted samples, manual correction of English fragments, and quality inspection by two native Urdu speakers on randomly selected samples. The native speakers reviewed a randomly selected subset of 500 samples (1% of the translated dataset), assessing fluency, semantic fidelity, and correctness of climate-specific terminology. Minor edits were made based on their feedback. The same translation pipeline was used consistently for both training and test sets.

1) Translation method: An entirely automated LLM-based pipeline was used for the translation. Every English instruction-response pair of Clima500-Instruct was entered into ChatGPT-4 with explicit prompts to preserve semantic meaning, technical terminology, and conversational tone. Special consideration was given to maintenance of domain specific vocabulary relevant to climate science (greenhouse gases,

deforestation, sea-level rise) and the prevention of literal, word-for-word translations that could distort meaning in Urdu.

2) Quality assurance and validation: To ensure the quality and usability of the ChatGPT-4 translated Urdu dataset, a systematic post-processing and manual filtering procedure was used. First, all the translated entries were automatically scanned in order to identify and remove malformed tokens, undefined symbols, and formatting artifacts introduced during generation. Second, remaining English words or phrases in Urdu questions or answers were manually detected and translated to Urdu to maintain linguistic consistency.

Third, the corrupted samples, where the response or the instruction was partially or fully in English, were eliminated. Lastly, to validate overall translation quality, two native Urdu speakers inspected, randomly selected subset of 500 samples (1% of the translated dataset), assessing fluency, semantic fidelity, and the correctness of climate-specific terminology. Their suggestions were then followed up with minor revisions for polishing and correcting the technical mistranslations in the downstream part of the phrasing. This post-processing pipeline guaranteed the final Urdu Clima500-Instruct subset was coherent, domain-faithful, and suitable for fine-tuning and retrieval augmented generation.

IV. Methodology

Fine-tuning is an important machine learning technique, especially in the field of natural language processing (NLP) [28]. This method involves a pre-trained model being further trained on a specialized dataset to adapt it for a specific task.Initially, large language models (LLMs) are trained on large amounts of general-purpose text data allowing them to develop a broad understanding of language. However, to obtain a higher degree of accuracy and contextual relevance in domain-specific tasks, fine-tuning is used.

In our work, we leverage fine-tuning to develop UrduClimateGPT, a model explicitly designed to produce retrieval-grounded and climate-related responses in the Urdu language. This was achieved by fine-tuning LLaMA 3.1 exclusively on the Urdu-translated 10% subset of the Clima500-Instruct dataset. This translated dataset plays a crucial role in fine tuning the model, allowing it to comprehend and verify claims in Urdu effectively. The fine-tuned Urdu-ClimateGPT model is a retrieval-grounded text generator, able to generate climate-focused responses based on external evidence under an automated evaluation protocol.

On top of fine-tuning, our approach incorporates retrieval augmented generation (RAG), a powerful technique that enhances contextual understanding and reliability of the generated responses. RAG combines the advantages of retrieval-based and generative models to reduce hallucination by grounding the generated text in verifiable and external sources of knowledge. In this method, relevant information is retrieved from a vector database, containing structured embeddings from authoritative sources such as Climate-FEVER. The retrieved context is then inputted into the language model, increasing the accuracy of its response and factual alignment.

A. System Architecture Overview

The overall architecture of Urdu-ClimateGPT follows a hybrid retrieval–generation design, integrating fine-tuning and retrieval augmented generation (RAG) to improve factual reliability and domain specificity. The system consists of four main components: 1) a preprocessing and embedding module, 2) a vector database for document storage, 3) a retrieval engine, and 4) a fine-tuned Urdu LLaMa model.

Climate-FEVER evidence sentences and translated Clima500-Instruct entries are embedded using a sentence transformer model and stored in the vector database. When a user query is received, semantically relevant documents are retrieved using similarity search and are appended as context to the fine-tuned Urdu-ClimateGPT model.

B. Vector Database Construction

To enable efficient semantic retrieval, we built a vector database containing dense embeddings based on two main sources, the Climate-FEVER dataset, and the translated Clima500-Instruct corpus. All pairs of evidence sentences and instructions-responses were embedded using a pre-trained sentence-transformer model.

The resulting vectors were indexed using a similarity search backend to handle quick approximate nearest-neighbor queries. Each of the vectors was accompanied by metadata such as source document ID, claim label (Support, Refute, Disputed), and topical category to facilitate filtered retrieval and structured grounding during inference.

C. Retrieval Augmented Generation Pipeline

Urdu-ClimateGPT uses a retrieval augmented generation (RAG) to ground responses in authoritative climate knowledge. After a user query, it first computes its embedding and retrieves the semantically-closest top-k entries of the vector database.

The documents returned are formatted into a structured context block appended with the original query using a predefined prompt template. The resulting enriched prompt is then passed to the fine-tuned Urdu-ClimateGPT model, which allows generating contextually aware and fact-grounded responses.

In our implementation, we set $k = 5$ to balance retrieval diversity and prompt length constraints. This design reduces hallucinations and improves factual consistency, particularly for climate claim verification tasks.

D. Claim Detection and Routing Logic

To differentiate factual claims from general informational queries, we implemented a lightweight claim detection mechanism using a general-purpose LLM (Base LLaMa 3.1). The claim detection model was not fine-tuned on the Urdu climate dataset and operates independently of the response generation pipeline to avoid training leakage or bias. Each user query is first classified into one of two categories: 1) factual claim or 2) informational inquiry.

If the query is classified as a factual claim, the system triggers the retrieval and verification pipeline, retrieving

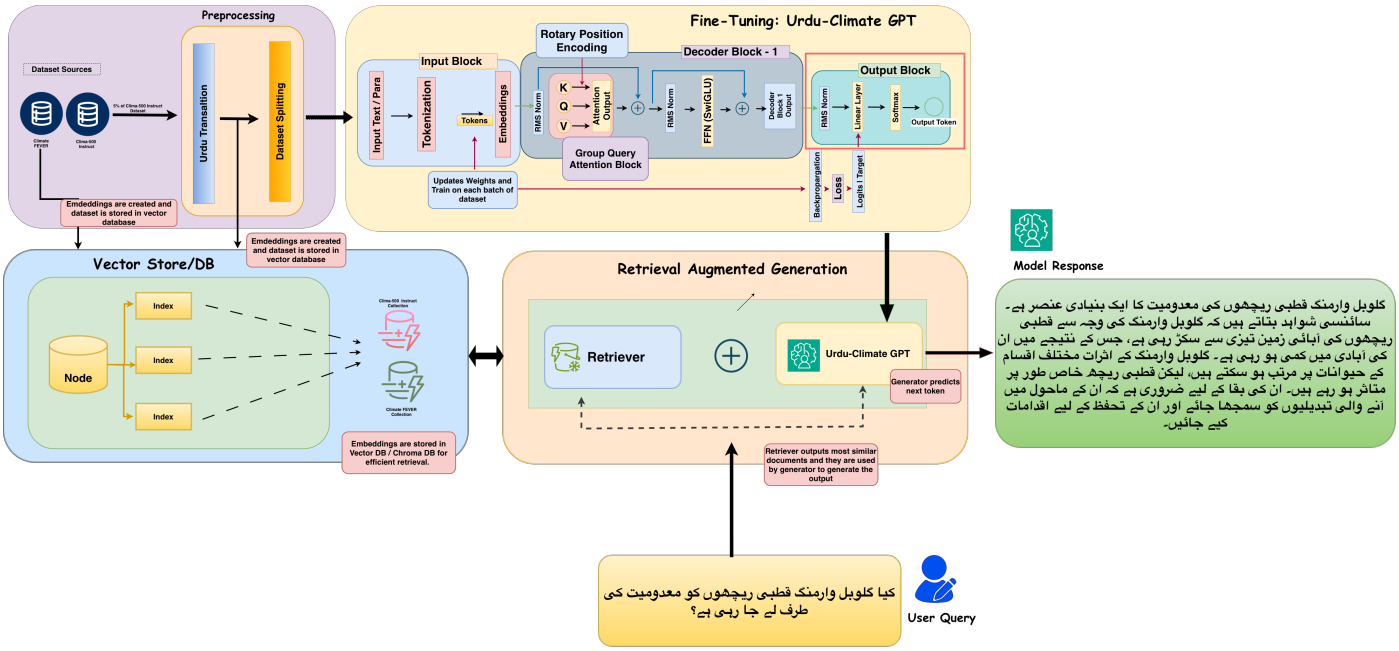


Fig. 4. Overview of the Urdu-ClimateGPT architecture integrating fine-tuning and retrieval augmented generation (RAG). Climate-FEVER embeddings are stored in a vector database for factual claim verification, while translated Clima500-Instruct embeddings are used for contextual grounding.

supporting or refuting evidence from the Climate-FEVER vector database. If the query is categorized as an informational inquiry, it is directly forwarded to the fine-tuned Urdu-ClimateGPT model.

This routing facilitates the optimal use of computational resources while maintaining high factual accuracy of claim-based queries.

The hybrid methodology provides factual credibility to the climate-related responses in the Urdu language, diminishes the hallucination-like effect, improves factual consistency, and ensures contextual accuracy in verification tasks. The overall methodology is illustrated in Fig. 4.

E. Embedding Model and Similarity Search

To generate dense vector representations for semantic retrieval, the pretrained sentence-transformer model `all-MiniLM-L6-v2` was used to encode Climate-FEVER evidence sentences and translated Clim500-Instruct entries. This model was chosen because of high quality semantic representation and computational efficiency [29], therefore making it suitable for real-time retrieval in a low resource deployment environment.

All text inputs were normalised before being embedded to ensure consistent vector representations across datasets. The resulting embeddings were saved in a vector database for near instant ANNS (approximate nearest-neighbor search).

The cosine similarity between the user query embedding and stored document embeddings was used for semantic retrieval. The top-k most similar entries were retrieved with a k value of 5 as it was found to be empirically validated for a balance between this retrieval diversity and the constraint of prompt length.

F. Vector Database Backend

To build the retrieval backend, ChromaDB, a high-dimensional similarity search vector store was used. All embedded documents were stored, each having metadata fields [source dataset (Climate-FEVER), claim label, topic category].

This metadata-based filtering allowed for the specific retrieval of claims verification queries and thus factual grounding to be based on Climate-FEVER evidence. The vector store has cosine similarity search and approximate nearest neighbor indexing to enable low-latency retrieval in real-time interactions.

G. Prompt Engineering for Retrieval Augmented Generation

Evidence passages were retrieved and inserted into the generation prompt through a template that consists of three parts: 1) system instructions to enforce factual grounding, 2) a contextual evidence block with the top-k retrieved passages and 3) the original user query.

The specific instructions to the model prompted it to stick to the evidence and refrain from speculative or unsubstantiated claims. The factual consistency of the generated statements and the avoidance of hallucination were improved with this controlled prompting strategy for climate claim verification.

H. Computational Efficiency and Latency

The retrieval and verification components introduce modest computational overhead compared to direct generation. On average, the full pipeline required approximately 1.4× the inference time of the base LLaMa 3.1 model due to embedding lookup and vector search. Despite this overhead, the system remained suitable for near real-time interaction in practical deployment scenarios.

V. Training Configuration

Key dataset, training, and hyperparameter configurations are summarized in Table I and Table II. We utilized a stratified 10% subset (50,000 samples) of the Clima500-Instruct (English) dataset, split into 49,800 training and 200 test prompts for fine-tuning Urdu-ClimateGPT. Fine-tuning was performed over 3 epochs on a Google Colab Pro GPU, requiring approximately 5 hours. Low-Rank Adaptation (LoRA) was applied with rank 16, scaling factor 32, and dropout 0.1, alongside AdamW optimization at a learning rate of 2×10^{-5} and a batch size of 8.

TABLE I. LoRA and Optimizer Hyperparameters

Parameter	Value
LoRA rank (r)	16
LoRA scaling (α)	32
LoRA dropout	0.1
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	8
Quantization	4-bit (bnb)

TABLE II. Dataset and Training Configuration

Parameter	Value
Dataset	Urdu-translated Clima500-Instruct
Total entries used	50,000
Sampling method	Stratified random sampling
Training set	49,800 entries
Test set	200 entries
Epochs	3
Training time	$\approx$5 hours
Compute environment	Google Colab Pro GPU

VI. Experiments and Results

To evaluate the effectiveness of our fine-tuned LLaMa model on Urdu climate change data, we conducted a pairwise comparison experiment against the baseline LLaMa model using ChatGPT-4 as a blinded LLM-based preference assessor. We curated a test set of climate change-related prompts in Urdu that were not seen during training. For each prompt, we generated responses from both the baseline and our fine-tuned model.

To assess qualitative performance, we employed a blind evaluation approach using a separate, unbiased chatbot-based preference assessor [30]. The two anonymized outputs (in randomized order) for each test prompt were presented to the assessor, who was asked to choose the better response or indicate if neither was preferable. This process is mirrored by the evaluation protocol shown in similar comparative studies, though it is adapted for Urdu and the climate change domain.

The preference percentages are presented in Table III. Our fine-tuned model was significantly preferred over the baseline, indicating strong performance improvements in generating relevant and fluent Urdu responses.

These results highlight the effectiveness of domain- and language-specific fine-tuning in improving response quality for specialized, underrepresented languages like Urdu.

TABLE III. Preference Results Comparing Our Fine-Tuned LLaMa-Urdu Model with the Baseline LLaMa on Urdu Climate Change Prompts.

Model	Ours	Competitor	Neither
Baseline Lama3.1	86.50%	12.00%	3%

A. Component-Wise Performance Comparison

The incremental impact of fine-tuning and retrieval augmented generation (RAG) on response quality was analyzed through a component-wise performance evaluation conducted across three system variants: 1) the baseline LLaMa 3.1 model, 2) the fine-tuned Urdu LLaMa model without RAG, and 3) the full Urdu-ClimateGPT system with both fine-tuning and RAG enabled.

The same held-out test set of 200 Urdu climate-related prompts was used to evaluate each model. The 200 held-out test prompts were obtained via stratified random sampling from the Clima500-Instruct dataset to preserve the original topic distribution. The resulting topic-wise distribution in the test set is as follows:

- Greenhouse Emissions: 57 prompts
- Climate Policy / Laws: 38 prompts
- Oceanic: 15 prompts
- Extreme Weather: 9 prompts
- Temperature: 6 prompts
- Precipitation: 5 prompts
- Land Cover, Hydrology, Air Quality, Renewable Energy: 70 prompts combined

This provides reasonable coverage across major climate topics aligned with the source dataset 3. For every prompt, scores were assigned along four qualitative dimensions: linguistic fluency, domain accuracy, factual consistency, and response completeness. The scores were normalized to a 0–1 scale and averaged across all test samples.

Fluency measures grammatical correctness and naturalness of Urdu output. Factual correctness and alignment with established climate knowledge are reflected by accuracy. Whether the response fully addresses the user query without missing key points is captured by completeness.

Fig. 5 illustrates the comparative performance of the three model variants. Limited accuracy and incomplete responses are shown by the baseline model. Fluency and domain accuracy are substantially improved by fine-tuning, while factual grounding and completeness are further enhanced by the addition of RAG. The highest overall performance across all four metrics is achieved by the full Urdu-ClimateGPT system, demonstrating that complementary benefits are contributed by fine-tuning and RAG.

VII. Ablation Study

To quantify the individual contributions of fine-tuning and retrieval augmented generation (RAG) in Urdu-ClimateGPT, an ablation study was conducted by systematically disabling

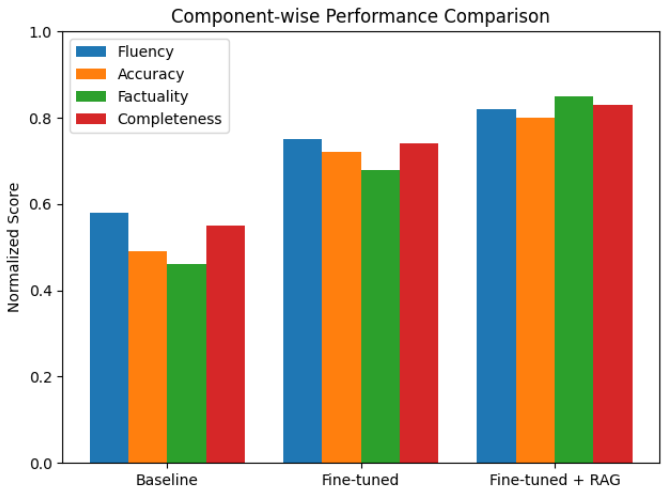


Fig. 5. Component-wise performance comparison of baseline LLaMa 3.1, fine-tuned Urdu LLaMa, and full Urdu-ClimateGPT across fluency, accuracy, factuality and completeness.

key components of the proposed system. Four configurations were evaluated: 1) the baseline LLaMa 3.1 model without any adaptation, 2) LLaMa 3.1 that was fine-tuned on the Urdu Clima500-Instruct dataset without RAG, 3) the baseline LLaMa 3.1 with RAG enabled but without fine-tuning, and 4) the full Urdu-ClimateGPT system with both fine-tuning and RAG enabled. The same held-out Urdu climate prompt set described in Section VI was used to test all configurations. Performance was measured by four criteria: linguistic fluency, domain accuracy, factual consistency, and response completeness. Each metric was normalized to a 0–1 scale and averaged across 200 evaluation prompts.

Table IV reports the comparative results. The baseline model performs weakest across all metrics, indicating limited capability in handling Urdu climate-specific queries. Fine-tuning alone yields substantial improvements in fluency and domain accuracy, while RAG alone enhances factual consistency by grounding responses in external climate evidence, while yielding moderate gains in domain accuracy and response completeness relative to the baseline. The full Urdu-ClimateGPT system achieves the highest overall score, demonstrating that fine-tuning and RAG provide complementary benefits and jointly reduce hallucinations while improving response quality.

TABLE IV. ABLATION STUDY ACROSS FLUENCY, ACCURACY, FACTUALITY, AND COMPLETENESS.

Model Variant	Fluency	Accuracy	Factuality	Completeness	Overall
Baseline LLaMa 3.1	0.58	0.49	0.46	0.55	0.52
+ Fine-tuning only	0.75	0.72	0.68	0.74	0.72
+ RAG only	0.64	0.63	0.76	0.65	0.67
Urdu-ClimateGPT (Full)	0.82	0.80	0.85	0.83	0.82

A. Urdu-ClimateGPT response vs Baseline LLaMa

Example 1: Fig. 6

Prompt: The impact of rising temperatures on the Earth's surface due to climate change.



Fig. 6. Urdu question on impact of rising temperature on sea level with model response.

Response 1 (Urdu-ClimateGPT): When temperatures rise on Earth due to climate change, many environmental and natural systems are affected. Climate change is caused by an increase in greenhouse gases in the atmosphere, which trap more heat. As a result, more heat is added to the Earth's surface, leading to several consequences. This not only increases the average temperature but also intensifies heatwaves. Due to rising temperatures, glaciers and polar ice are melting, which causes sea levels to rise. This increases the risk of flooding in coastal areas. In addition, weather patterns are changing, leading to extreme events such as droughts, floods, and storms. Ecosystems are also affected by rising temperatures. Difficulties are faced by many plants and animals in adapting to new conditions, which threatens biodiversity. The productivity of agriculture is affected, as crops could fail due to high temperatures and lack of water. Further, increases in temperature, which will reduce the availability of clean drinking water, can contribute to water scarcity. In conclusion, the rising temperature of the earth's surface as a result of climate change is a serious issue that is impacting human life, ecosystems and natural resources. Global actions like greenhouse gas reduction and sustainable practice are needed to mitigate these impacts.

Response 2 (LLaMa 3.1): When temperatures rise on Earth due to climate change, it leads to many major environmental changes. One major effect is that ice is melting in the poles causing sea levels to rise. This poses a risk of flooding for coastal cities and island countries. Furthermore, elevated temperatures makes extreme weather events, like heatwaves, droughts, hurricanes and heavy rainfall more likely and intense. They inflict damage to infrastructure, losses to agriculture and endanger human lives. Ecosystems are also impacted by the increase of temperature. Many species are struggled with to survive as their natural habitats are changed or disappeared. Coral reefs suffer from bleaching caused by warmer ocean temperatures, and forest are increased in their susceptibility to fire. Another major impact that is recognized as serious is the scarcity of water. The evaporation rates are increased with rising temperature, which decreases the availability of freshwater. This poses challenges for drinking water supply, irrigation and energy production. Overall, widespread and long-term impacts on the Earth's surface, human societies, and natural systems are had by climate change–driven temper-

ature increases. This issue demands international collaboration, carbon emissions reduction, and a transition to renewable energy.

Example 2: Fig. 7



Fig. 7. Urdu question on carbon dioxide and tree planting with model responses.

Prompt: Reducing carbon dioxide in the atmosphere by planting trees.

Response 1 (Urdu-ClimateGPT): Trees play a vital role in reducing carbon dioxide because they absorb CO_2 during photosynthesis and release oxygen. When more trees are planted, more carbon dioxide is removed from the air, helping to reduce greenhouse gas concentrations. This process slows global warming and improves air quality. Trees also help regulate temperatures by providing shade and releasing moisture into the atmosphere, which cools the surrounding environment. In addition, forests support biodiversity by providing habitats for animals and plants. Planting trees also improves soil quality, prevents erosion, and enhances water retention. Overall, increasing tree cover is a natural and effective way to mitigate climate change, protect ecosystems, and promote environmental sustainability.

Response 2 (LLaMa 3.1): Planting trees can help reduce carbon dioxide by absorbing CO_2 from the atmosphere through photosynthesis, but it is not a complete solution to climate change. Time is taken by trees to grow, and the carbon absorption capacity is limited. In some regions, large-scale tree planting may not be deemed suitable due to water scarcity or land-use conflicts. Forests can also be lost to wildfires, pests, or deforestation, with stored carbon being released back into the atmosphere. Tree planting is seen as helpful but it should be coupled with the phasing out of fossil fuel utilization, the safeguarding of existing forests and switching to renewable energy.

Example 3: Fig. 8

Prompt: Discuss the causes of air pollution in cities and suggest solutions.

Response 1 (Urdu-ClimateGPT): Urban air pollution is a serious problem affecting both the environment and human health. It is caused by different factors such as vehicle emissions, industrial activities, construction dust, fossil fuel combustion and inadequate waste management. The deterioration



Fig. 8. Urdu question on urban air pollution and model responses.

of air quality is due to the emission of polluting gases like carbon monoxide, nitrogen oxides, sulfur dioxide and particulate matter (PM2.5). Exposure to polluted air over a long period of time increases the risk of respiratory diseases, heart disease, and premature death. Air pollution makes it difficult to see, damages buildings and affects urban ecosystems. In response to this, it is necessary to encourage public transport, the use of clean energy, environmental laws, the creation of green areas and sensitizing the public. Good policies and community engagement play a vital role in the process of improving air quality and thereby creating better living conditions in urban areas.

Response 2 (LLaMa 3.1): In cities, traffic congestion, industrial emissions, and the use of fossil fuels are the three main factors contributing to air pollution. Bad urban planning and high population density due to the growing number of vehicles and energy needs make the situation worse. Air pollution also has a negative impact on human health, contributing to asthma, lung infections and other respiratory diseases. Climate change and environmental degradation are also contributed to by it. This problem needs to be addressed by coordinated efforts of governments, industries and citizens. Improving public transport systems, promoting the use of electric cars, controlling industrial pollution, and planting more trees in urban areas are some of the measures that can help to reduce air pollution. In the end, mitigating urban air pollution is a public health issue as well as the environment.

VIII. Limitations

Urdu-ClimateGPT is a fine-tuned language model based on LLaMa architecture. Various data, such as large-scale web and publicly available text data, were used for pretraining the base model. Although a wide language functionality is given as a benefit by this, the risk of inaccuracy with the training data is present.

We employed a vector-based retrieval augmented generation (RAG) database to check the factuality of model-generated claims to enhance factual correctness and context relevance. These parts help considerably to improve the credibility of outputs, especially with regard to the use of climate discussion. There are, however, a few issues with these safeguards. The bias behaviour, however, can be continued in the underlying language model and/or the model can give over-confident and

speculative and subtly wrong answers [31]. These problems are particularly probable in case of insufficient context, ambiguous prompts or where the retrieval and judgment systems incorrectly categorize the query.

Therefore, users and researchers are advised to take Urdu-ClimateGPT's answers with a pinch of salt and think critically. The results of the models are not considered to be definitive until they are externally verified. It is seen as important to acknowledge these limitations for AI systems to be deployed more responsibly and to continue to be developed in the most sensitive and high-impact areas, including climate communication. Further, it is noted that automated evaluation may not be sufficient to cover the factual reliability in the real world, which is a major focus of our human expert validation.

IX. Ethics Statement

The research is entirely based on publicly accessible and open source data which does not contain any personal or identifiable data hence there is no privacy or ethical risk. All contributions and ideas are original and reference to all the datasets and models used has been made. It is important to acknowledge that, similar to any generative language model, the system can generate dangerous or offensive results when used irresponsibly and therefore, responsible use and careful monitoring during deployment are crucial.

X. Future Work

While this research uses LLM-based automated evaluation for scalability and consistency, the critical next step is the introduction of the structured evaluation by human experts to ensure factual correctness and real-world utility. In future work, we plan to use a multi-rater human assessment including climate science graduate students, environmental researchers, and Urdu-language educators. The human evaluation protocol will be used to evaluate model outputs in four dimensions, including: 1) linguistic fluency and grammatical correctness in Urdu, 2) factual accuracy in relation to established climate science, 3) evidence consistency with retrieved sources, and 4) response completeness and usefulness to end users. Each response will be independently rated by at least three annotators using a five-point Likert scale, with inter-annotator agreement measured using Krippendorff's alpha. The evaluators will not be told for any given response, which system variant produced it, to enable methodological rigor. The disagreements will be solved by means of majority voting and expert adjudication. In addition, a set of questions will be developed that will test extreme edge cases of ambiguous assertions, misinformation elicitation and culturally-relevant climate scenarios relevant to Urdu-speaking communities. This human-centered assessment will offer authoritative confirmation of factual accuracy, complement automated metrics, and establish a strong empirical foundation of Urdu-ClimateGPT implementation in educational, journalistic, and policy-support practice.

References

[1] S. C. Moser, "Communicating climate change: History, challenges, process and future directions," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 1, no. 1, pp. 31–53, 2010.

[2] K. Nissen and H. Wong, "Democratizing climate change knowledge: The role of digital platforms," *Environmental Communication*, vol. 13, no. 3, pp. 345–360, 2019.

[3] Ethnologue, "What are the top 200 most spoken languages?" 2025, accessed: May 2025. [Online]. Available: https://www.ethnologue.com/insights/ethnologue200/

[4] "25 most spoken languages in the world in 2025," https://www.berlitz.com/blog/most-spoken-languages-world, 2024, accessed: 2025.

[5] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the nlp world," *Proceedings of ACL*, 2020.

[6] UNESCO, "Urdo — silk roads programme – unesco language profile," https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/urdo, 2025, accessed: 2025.

[7] OpenAI, "Chatgpt," 2024, accessed: 2025-02-18. [Online]. Available: https://openai.com/index/chatgpt/

[8] Anthropic, "Claude ai," 2024, accessed: 2025-02-18. [Online]. Available: https://claude.ai

[9] Google, "Gemini ai," 2024, accessed: 2025-02-18. [Online]. Available: https://gemini.google.com/

[10] S. S. Mullappilly, A. Shaker, O. Thawakar *et al.*, "Arabic mini-climategpt: A climate change and sustainability tailored arabic llm," *arXiv preprint*, no. 2312.09366, 2023. [Online]. Available: https://arxiv.org/abs/2312.09366

[11] D. Thulke, Y. Gao, P. Pelser, R. Brune, R. Jalota, F. Fok, M. Ramos, I. van Wyk, A. Nasir, H. Goldstein, T. Tragemann, K. Nguyen, A. Fowler, A. Stanco, J. Gabriel, J. Taylor, D. Moro, E. Tsymbalov, J. de Waal, E. Matusov, M. Yaghi, M. Shihadah, H. Ney, C. Dugast, J. Dotan, and D. Erasmus, "Climategpt: Towards ai synthesizing interdisciplinary research on climate change," *arXiv preprint*, no. 2401.09646, 2024. [Online]. Available: https://arxiv.org/abs/2401.09646

[12] Y. Z. Md Rashad and Others, "Climatebot: A machine reading comprehension bot for climate change documents," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 5243–5250. [Online]. Available: https://www.ijcai.org/proceedings/2022/729

[13] T. Laud, D. Spokoyny, T. Corringham, and T. Berg-Kirkpatrick, "Climabench: A benchmark dataset for climate change text understanding in english," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2301.04253

[14] T. Diggelmann *et al.*, "Climate-fever: A dataset for verification of real-world climate claims," *arXiv preprint*, no. 2012.00614, 2020.

[15] A. Biswas, H. Chahout, T. Pigram, H. Dong, H. T. Williams, F. Fung, and H. Xie, "Evaluating retrieval augmented generation to communicate UK climate change information," in *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*. Association for Computational Linguistics, 2025, pp. 126–141. [Online]. Available: https://aclanthology.org/2025.climatenlp-1.9/

[16] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," *Findings of ACL*, 2021.

[17] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[18] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint*, no. 2007.01282, 2020.

[19] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 6769–6781.

[20] O. Ram *et al.*, "In-context retrieval-augmented language models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316–1331, 2023.

[21] L. Huayang *et al.*, "A survey on retrieval-augmented text generation," *arXiv preprint*, no. 2202.01110, 2022.

[22] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint*, no. 2312.10997, 2023.

[23] S. Arif *et al.*, "Evaluating large language models for urdu," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2407.04459

[24] F. Adeeba, B. Dillon, H. Sajjad, and R. Bhatt, "Urblimp: A benchmark for evaluating the linguistic competence of large language models in urdu," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2508.01006

[25] A. Hashmat, M. A. Mirza, and A. A. Raza, "Pakbbq: A culturally adapted bias benchmark for qa," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2508.10186

[26] W. Jiao, X. Wang, S. Zhang, Y. Li, Y. Zhang, and Z. Tu, "Is chatgpt a good translator? a preliminary study," *arXiv preprint arXiv:2301.08745*, 2023.

[27] L. Wang, Z. Tu, A. Way, and Q. Liu, "Large language models are state-of-the-art translators," *Transactions of the ACL*, 2023.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, 2020.

[29] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *EMNLP*, 2019.

[30] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Y. Wu, H. Zhuang, Z. Lin, Z. Zhang, J. Gonzalez, and I. Stoica, "Judging llm-as-a-judge: Bias and limitations of llm-based evaluation," *NeurIPS*, 2023.

[31] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots," *Proceedings of FAccT*, 2021.

Leveraging MCESTA and Large Language Models for Next-Generation Similarity Learning

Mohamedou Cheikh Tourad[1]*, Abdelmounaim Abdali[2], Mohamed Dhleima[3],
Naoual Mouhni[4], Sana Chakri[5], Ibtissam Amalou[6], Saadbouh Cheikh El Mehdy[7]
Faculty of Sciences and Techniques, University of Nouakchott, Al-Khawarizmi, Nouakchott, Mauritania[1,3,7]
Laboratory of Mathematics, Artificial Intelligence and Sustainable Technologies, Cadi Ayyad University, Marrakech, Morocco[2]
GL-ISI Team, Department of Computer Science, FSTE, Moulay Ismail University (UMI), Meknès, Morocco[4]
LIM Laboratory, Hassan II University of Casablanca, Mohammedia, Morocco[5]
LAMIGEP, Moroccan School of Engineering Sciences (EMSI), Marrakech, Morocco[6]

***Abstract*—Large Language Models (LLMs) have reshaped how machines read and compare text, yet most similarity learning pipelines built on top of them still behave like black boxes: a single cosine score is returned without any indication of *why* two sentences were deemed close. This study proposes a different route. We pair a fine-tuned Sentence-BERT (SBERT) encoder with MCESTA, a fuzzy multi-criteria aggregation layer that combines a semantic (cosine), a geometric (Manhattan) and a lexical (Jaccard) similarity through a small set of human-readable linguistic rules. The output is a single similarity score in $[0, 1]$ that remains traceable to the rules that produced it: because the rule base contains only twelve Mamdani rules, the chain of fired antecedents can be inspected directly after each inference, which is what we mean by *traceability* in this study. We evaluate the framework on the Quora Question Pairs corpus (the public release contains about $404{,}000$ question pairs in English with a $63/37$ non-paraphrase to paraphrase split) against five strong baselines, including SimCSE and AnglE. Our model reaches Accuracy $= 0.90$ and AUC $= 0.94$, outperforming every baseline. A controlled ablation shows that the gains come from the fuzzy aggregation step itself, not from the choice of encoder, while a robustness study reveals that the soft membership functions absorb noise and threshold variations more gracefully than a plain cosine baseline. The fuzzy aggregation step runs in $\mathcal{O}(|R|)$ per pair, where $|R| = 12$ is the size of the rule base, so its computational overhead on top of the encoder forward pass is negligible. Adaptive fuzzy rules, multilingual similarity, and domain-specific deployments are positioned as future extensions rather than as results of the present study.**

***Keywords*—Similarity learning; MCESTA; fuzzy aggregation; Natural Language Processing; Large Language Models; Sentence-BERT; explainable AI**

I. Introduction

Measuring how close two pieces of text really are is one of those problems that looks easy at first sight and turns out to be surprisingly subtle. Search engines, recommendation systems, paraphrase detectors, plagiarism scanners, and clinical decision tools all rely on a similarity score under the hood [1], [2]. In Natural Language Processing (NLP), this is usually phrased as *semantic textual similarity*: given two sentences, decide how close they are in meaning, ignoring whatever surface differences happen to be there [2], [3].

For a long time, the dominant approach was lexical: count overlapping words, weight them with TF-IDF, run BM25, compute a Jaccard coefficient. These tricks work reasonably well for keyword-heavy queries, but they break as soon as the two sentences share little vocabulary while saying the same thing, and they are easily fooled by negation or paraphrase [1]. The introduction of transformer-based Large Language Models (LLMs) [4], [5] changed the picture almost overnight. By projecting each sentence into a dense contextual embedding, models such as Sentence-BERT [6], [7], SimCSE [8], AnglE [9] and the more recent contrastive E5 family [10] place semantically related sentences close to one another in vector space, and currently sit at the top of leaderboards such as the Massive Text Embedding Benchmark [11] and the Quora Question Pairs corpus [12], [13].

Yet two issues keep coming back. The first one is that a single cosine value collapses very different kinds of evidence—meaning, shared vocabulary, geometric proximity in embedding space—into one number, and that number can be hard to defend when the sentences are similar along one axis but different along another [14]. The second issue is more practical: when an end-user (a doctor, a lawyer, a content moderator) asks *why* two sentences were declared equivalent, the model has nothing to show beyond an opaque scalar. This is a real obstacle in regulated domains [15]–[17].

Fuzzy aggregation offers a useful counterweight. Instead of collapsing several similarity signals into a single learned scalar, a fuzzy inference system aggregates them through a small set of linguistic rules that anyone can read [14], [18]. MCESTA [19], [20], the framework we build upon in this study, was designed exactly with this trade-off in mind: it combines several heterogeneous similarity measures through a Mamdani-type controller whose weights are not back-propagated, but inferred from human-readable rules.

A. Motivation

This study sits between these two worlds. On the one side, LLM embeddings give us semantic richness; on the other side, fuzzy multi-criteria aggregation gives us transparency and a degree of robustness that purely neural fusion layers do not provide. The literature has explored each direction separately, but combining them in a clean, reproducible pipeline is far less common. We chose the Quora Question Pairs (QQP) benchmark [12], [21] for this study for three concrete reasons. First, it is large enough that small differences between aggregation strategies become statistically meaningful. Second, the

*Corresponding author

mild class imbalance between paraphrase and non-paraphrase pairs stresses the calibration of any similarity score. Third, it is the de facto benchmark used by SBERT [6], SimCSE [8] and AnglE [9], which makes a head-to-head comparison straightforward.

B. Problem Statement and Research Questions

The problem we address can be stated compactly. Given a pair of short texts (S_1, S_2), the goal is to predict a binary paraphrase label $\hat{y} \in \{0, 1\}$ together with a similarity score in $[0, 1]$ that downstream users can audit. Three research questions guide the rest of this study.

RQ1. Can a non-differentiable, rule-based aggregation of three complementary similarity measures match or exceed the accuracy of strong neural baselines such as SimCSE and AnglE on QQP?

RQ2. How much of the observed accuracy gain is attributable to the fuzzy aggregation *strategy* versus the underlying encoder?

RQ3. Does the fuzzy aggregation step improve the calibration and robustness of the resulting similarity scores compared with a plain cosine pipeline?

C. Contributions

The main contributions of this study can be summarised as follows:

- We design a hybrid pipeline that plugs a fine-tuned SBERT encoder into the MCESTA fuzzy aggregation layer, fusing *semantic* (cosine), *lexical* (Jaccard) and *geometric* (Manhattan) signals through an explainable Mamdani fuzzy inference system.
- We benchmark the framework on QQP [13], [21] against five strong baselines (LSTM Siamese, SBERT+Cosine, SBERT+Weighted Sum, SimCSE [8], and AnglE [9]). Our model reaches Accuracy $= 0.90$ and AUC $= 0.94$, ranking first on every reported metric.
- We complement the headline numbers with a deeper analysis: training-loss curves, ROC and precision-recall plots, a confusion-matrix breakdown, two ablation studies on the aggregation layer and on each individual similarity component, and a robustness study under rule perturbation, Gaussian noise and threshold variation.
- Finally, we discuss what the fuzzy rule base actually buys in practice—in particular how easy it is to inspect, audit and extend—and we sketch how the framework can be grown towards neuro-fuzzy [18], multilingual [22], and domain-specific similarity tasks.

D. Paper Organisation

The rest of the study is organised as follows: Section II reviews the background that the framework rests on: similarity learning, LLMs, and the MCESTA fuzzy aggregation method. Section III describes the proposed methodology end-to-end. Section IV reports the experiments and the ablation/robustness studies. Section V draws the main conclusions, openly discusses what the current framework cannot yet do, and lists the research questions we plan to tackle next.

II. Background

A. Similarity Learning

At its core, similarity learning is about teaching a model to tell how close two data items are [23]. Distance metrics such as Euclidean and cosine similarity are the historical workhorses, and they remain useful baselines today; deep variants such as Siamese networks with triplet loss [24] go one step further by *learning* an embedding space in which the geometric distance already reflects semantic closeness. A Siamese network is essentially a pair of twin subnetworks with tied weights, fed two inputs in parallel and asked to produce comparable feature vectors.

In practice, a Siamese pipeline has two ingredients:

- Feature extraction. Each branch maps the input to a dense vector. The encoder can be a GRU, an LSTM, or a pre-trained transformer; what matters is that both branches share weights so that semantically similar inputs end up close to each other in the embedding space [25].
- Distance computation. The two output vectors are compared with a metric—most often cosine or Euclidean. Smaller distances indicate greater similarity.

B. Contrastive Loss vs. Triplet Loss

1) Contrastive Loss: Contrastive loss [26] works on pairs. The idea is intuitive: similar pairs should be pulled together, dissimilar pairs should be pushed at least a margin apart. Given inputs x_1, x_2, embeddings $f(x_1), f(x_2)$, and a label $y \in \{0, 1\}$ that flags whether the pair is similar, the loss reads:

$$L = (1 - y) \cdot \tfrac{1}{2} d^2 + y \cdot \tfrac{1}{2} \max(0, m - d)^2 \tag{1}$$

with $d = \|f(x_1) - f(x_2)\|_2$ and a margin m. Summing this loss over all training pairs gives the global objective $J = \sum_{i=1}^{n} L(x_1^{(i)}, x_2^{(i)}, y^{(i)})$ that the network minimises [27].

2) Triplet Loss: Triplet loss is the variant most often used for face recognition and image retrieval. Each training example is built around three items—an anchor A, a positive P (same class as the anchor), and a negative N (different class)—and the loss:

$$\mathcal{L}_{\text{trip}} = \max\big(d(A, P) - d(A, N) + \alpha,\ 0\big) \tag{2}$$

forces the anchor to be at least α closer to the positive than to the negative [20].

C. Large Language Models (LLMs)

LLMs [5], [6] are deep transformer networks pre-trained on huge text corpora. The output of a single forward pass is a contextual embedding—a vector that summarises the meaning of an input in light of its surrounding context. Models in the BERT, GPT and RoBERTa families excel on tasks that require this kind of semantic understanding, and serve as the encoder of choice in modern sentence similarity pipelines [4].

D. MCESTA Method

Definition 1 (MCESTA)**.** *Let* $x, y \in \mathbb{R}^d$. *The* Multi-Component Enhanced Similarity and Thresholding Approach (MCESTA) *(Mohamedou Cheikh Elghotob Cheikh Saad bouh Cheikh Tourad Abass) aggregates* K *elementary similarity measures* $\{S_k\}_{k=1}^{K}$ *with adaptive weights inferred by a fuzzy inference system:*

$$\mathrm{MCESTA}(x, y) = \sum_{k=1}^{K} \alpha_k \, S_k(x, y), \quad \sum_{k=1}^{K} \alpha_k = 1 \tag{3}$$

where, $\alpha_k \in [0, 1]$ is the contribution of the k-th component. Weights are inferred through fuzzy rules evaluating the relevance of each measure [28].

In short, MCESTA blends several similarity measures by giving each one an adaptive weight that depends on the input [19]. A Mamdani-type fuzzy controller estimates these weights from the cosine and Jaccard scores [19], and the final aggregation takes the form:

$$\mathrm{MCESTA}(\widetilde{X}, \widetilde{Y}) = \sum_{i=1}^{p} \gamma_i \, R_i(\widetilde{X}, \widetilde{Y}) \tag{4}$$

$$R_i(\widetilde{X}, \widetilde{Y}) = \sum_{j=1}^{q} \delta_j \, R_{ij}(\widetilde{X}, \widetilde{Y}) \tag{5}$$

with $\sum_i \gamma_i \leq 1$ and $\sum_j \delta_j \leq 1$.

E. Positioning vs. Prior Hybrid and Explainable Similarity Frameworks

Hybrid similarity pipelines come in three main flavours. *Late-fusion linear* combinations (e.g., SBERT + Weighted Sum) learn a single scalar weight per similarity measure; they are simple but cannot express conditional patterns such as "medium cosine and high lexical overlap implies high similarity". *Neuro-fuzzy* hybrids [18] inject fuzzy logic inside a differentiable network; they recover some of the expressivity but lose the rule-by-rule auditability that motivates fuzzy systems in the first place. *Rule-based* fuzzy aggregators such as MCESTA [19], [20] sit at the other end of the spectrum: they keep the rule base fully readable and modular, at the cost of being non-differentiable. Table I summarises this landscape qualitatively.

What is therefore specific to our approach is the combination of (1) a contrastive sentence encoder fine-tuned with cross-entropy, (2) three complementary similarity signals (semantic, geometric, lexical), and (3) a small, fully auditable Mamdani rule base whose fired rules can be inspected at inference time.

III. Methodology

A. Overall Architecture

The framework we propose pairs a transformer encoder with an explainable multi-criteria aggregation layer (MCESTA). It is deliberately not end-to-end—we want the boundary between "learned representation" and "decision logic" to remain clearly drawn so that each stage stays inspectable.

Fig. 1 traces the data flow. Given an input pair (S_1, S_2), the SBERT encoder maps each sentence to a 768-dimensional vector, yielding the embeddings $(\mathbf{u}, \mathbf{v})$. Three similarity measures are then computed in parallel from these embeddings, each capturing a different facet of resemblance: Jaccard (lexical overlap), cosine (semantic alignment) and Manhattan distance through an exponential kernel (geometric proximity). MCESTA then aggregates these three signals using a Mamdani-type fuzzy inference system whose weights are derived from a small, human-readable rule base. The output is a single crisp score in $[0, 1]$, which is thresholded to produce the final binary paraphrase decision.

B. Sentence Embedding Extraction

Each sentence goes through a pre-trained SBERT encoder, which returns a fixed-length dense vector:

$$\mathbf{u} = \mathrm{Encoder}(S_1), \quad \mathbf{v} = \mathrm{Encoder}(S_2), \tag{6}$$

with $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. We fine-tune the encoder with a binary cross-entropy objective on the training split, and then freeze its parameters: the encoder is treated as a fixed feature extractor while the aggregation layer takes over.

C. Multi-Similarity Computation

We use three similarity measures because they capture three complementary, low-correlation views of the input pair: the cosine score measures *semantic alignment* in the learned embedding space, the Manhattan distance measures *geometric proximity* (which reacts to amplitude and dimension mismatches that cosine ignores), and the Jaccard coefficient measures *surface lexical overlap* computed directly on the raw tokens. The component-removal study reported in Section IV (see Table IV) shows that each of the three measures contributes a distinct, non-zero share to the final accuracy, which is why we did not add further measures in the current study.

1) Semantic similarity:

$$S_{\cos}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \tag{7}$$

2) Distance-based similarity:

$$S_{\mathrm{man}}(\mathbf{u}, \mathbf{v}) = \exp(-\|\mathbf{u} - \mathbf{v}\|_1) \tag{8}$$

TABLE I. QUALITATIVE POSITIONING OF MCESTA VS. REPRESENTATIVE HYBRID AND EXPLAINABLE SIMILARITY FRAMEWORKS. COLUMNS DESCRIBE DESIGN CHOICES, NOT MEASURED PERFORMANCE.

Framework	Encoder	Aggregation	Differentiable?	Rule-level audit?	Multi-criteria?
LSTM Siamese	GloVe + LSTM	cosine	yes	no	no
SBERT + Cosine [6]	SBERT	cosine	yes	no	no
SBERT + Weighted Sum	SBERT	learned linear	yes	no	yes
SimCSE [8]	contrastive SBERT	cosine	yes	no	no
AnglE [9]	angle-optimised SBERT	angle distance	yes	no	no
Neuro-fuzzy hybrids [18]	deep encoder	fuzzy + soft	yes	partial	yes
SBERT + MCESTA (Ours)	SBERT (fine-tuned)	Mamdani fuzzy rules	no	**yes**	**yes**

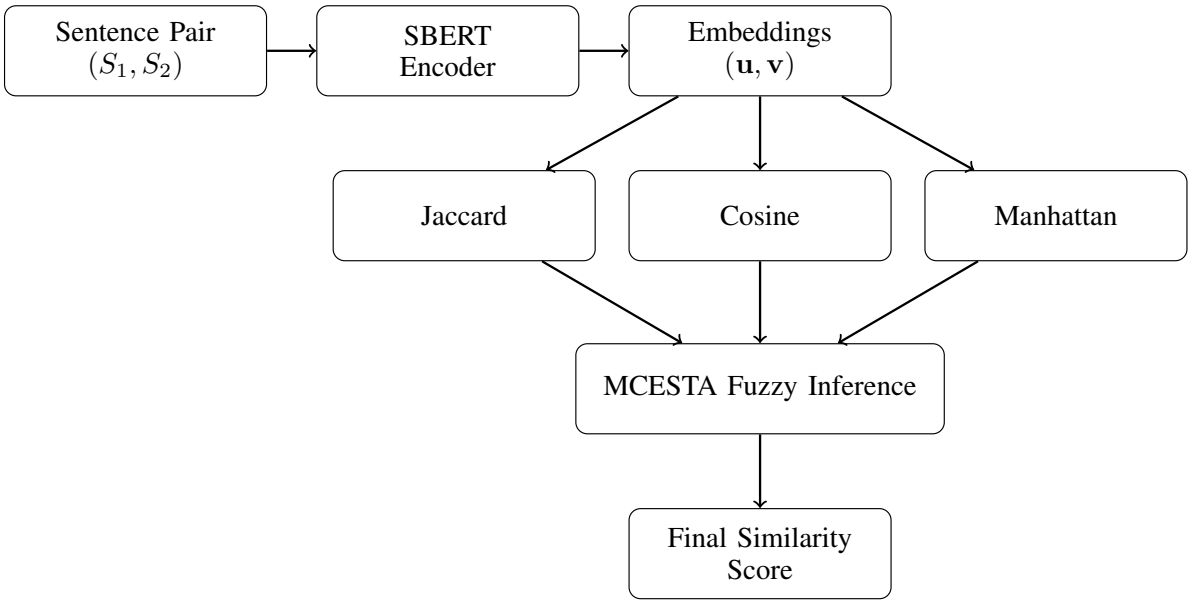


Fig. 1. Architecture of the proposed MCESTA-based similarity framework.

3) Lexical overlap similarity:

$$S_{\text{jac}}(S_1, S_2) = \frac{|T(S_1) \cap T(S_2)|}{|T(S_1) \cup T(S_2)|} \tag{9}$$

where, $T(\cdot)$ is the token set after normalization. All scores are normalized to $[0, 1]$.

D. MCESTA: Fuzzy Inference-Based Aggregation

1) Linguistic variables and membership functions: Each similarity measure is fuzzified into three linguistic levels—*Low*, *Medium*, and *High*—using triangular or trapezoidal membership functions. We deliberately keep the shapes simple: triangular and trapezoidal kernels are easy to read off a plot and easy to justify to a non-specialist. As an example, the cosine similarity is fuzzified as Low $[0.0, 0.0, 0.5]$, Medium $[0.3, 0.5, 0.7]$, and High $[0.6, 1.0, 1.0]$. A similar parametrisation is applied to S_{man} and S_{jac}.

2) Rule base design: The rule base is designed so that semantic similarity drives the decision while the lexical and distance signals refine it. Representative rules:

1) IF S_{cos} is High AND S_{jac} is High THEN High.
2) IF S_{cos} is High AND S_{man} is Medium THEN High.
3) IF S_{cos} is Medium AND S_{jac} is Medium THEN Medium.
4) IF S_{cos} is Low AND S_{jac} is Low THEN Low.
5) IF S_{cos} is Medium AND S_{jac} is Low THEN Medium.
6) IF S_{cos} is Low AND S_{man} is High THEN Medium.

3) Inference and defuzzification: We use a standard Mamdani configuration: each rule is activated through the minimum operator, the activated outputs are aggregated through the maximum operator, and the final crisp score is obtained by centroid defuzzification:

$$S_{\text{MCESTA}} = \frac{\int z\, \mu(z)\, dz}{\int \mu(z)\, dz} \tag{10}$$

E. Similarity-Based Classification

The continuous score $S_{\text{MCESTA}} \in [0, 1]$ is finally turned into a binary decision by comparing it to a threshold τ that we tune on the validation split:

$$\hat{y} = \begin{cases} 1, & S_{\text{MCESTA}} \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

F. Design Rationale

We deliberately keep the MCESTA layer non-differentiable. Plugging a fuzzy block into the gradient graph would force us to soften the rule logic and would erode the very property that motivates the whole design, namely traceability. By keeping the aggregation step rule-based and post-hoc, we keep interpretability and modularity, and the architecture stays in line with current explainable-AI principles [16], [17].

G. Inference Procedure (Pseudocode)

Algorithm 1 summarises the inference procedure for one sentence pair. Training follows the same pipeline with the encoder unfrozen and a binary cross-entropy loss back-propagated through the SBERT branch only; the MCESTA layer is kept frozen at all stages.

Algorithm 1. SBERT + MCESTA inference for a sentence pair.

Input: sentence pair (S_1, S_2), fine-tuned encoder $\text{Enc}(\cdot)$, fuzzy rule base R, threshold τ

Output: binary paraphrase decision $\hat{y}$ and crisp score S_{MCESTA}

$\mathbf{u} \leftarrow \text{Enc}(S_1)$ // d=768 embedding
$\mathbf{v} \leftarrow \text{Enc}(S_2)$
$s_{\cos} \leftarrow \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$
$s_{\text{man}} \leftarrow \exp(-\|\mathbf{u}-\mathbf{v}\|_1)$
$s_{\text{jac}} \leftarrow |T(S_1) \cap T(S_2)| / |T(S_1) \cup T(S_2)|$
Fuzzify each score into {Low, Medium, High}
Activate every rule $r \in R$ with the *min* operator
Aggregate activations with the *max* operator
$S_{\text{MCESTA}} \leftarrow \text{Centroid}(\mu_{\text{out}})$
return $\hat{y} = \mathbb{1}[\, S_{\text{MCESTA}} \geq \tau \,]$

H. Complexity Analysis

Let n be the number of tokens per sentence, $d = 768$ the embedding dimension, and $|R| = 12$ the rule-base size. The encoder forward pass dominates the cost at $\mathcal{O}(n^2 d)$ per pair, which is the standard transformer cost and is identical for all evaluated SBERT baselines. The three similarity scores are then computed in $\mathcal{O}(d)$ for cosine and Manhattan and in $\mathcal{O}(n)$ for Jaccard. The fuzzy aggregation is the only component specific to MCESTA: each rule is evaluated in constant time and the rule base is traversed once, giving an aggregation cost of $\mathcal{O}(|R|) = \mathcal{O}(12)$ per pair. In memory, MCESTA adds only the rule base itself on top of the encoder. The explainability overhead is therefore small relative to the encoder, both in time and in memory.

IV. Experiments and Results

A. Dataset and Experimental Protocol

All experiments are run on the Quora Question Pairs (QQP) dataset [21], a widely used benchmark for paraphrase identification [12]. The public QQP release contains 404,290 labelled question pairs in English with a mild class imbalance, approximately 63% non-paraphrases and 37% paraphrases [13], [21]. We split the corpus into 70%/10%/20% for training, validation, and test respectively, with the constraint that no question appears in more than one split—this avoids the leakage that QQP is notoriously prone to. We report Accuracy, Precision, Recall, F1, and AUC. The classification threshold τ is selected on the validation split by maximising F1, and the test split is touched only once for the final numbers. To reduce the influence of random initialisation, each run is repeated three times with seeds $\{42, 123, 2024\}$ and we report the mean across runs.

B. Baselines

We compare our framework against a representative mix of classical and recent baselines:

- LSTM Siamese – recurrent baseline with GloVe embeddings, included as a non-transformer reference point.
- SBERT + Cosine – SBERT embeddings compared with cosine similarity, the most common modern baseline.
- SBERT + Weighted Sum – the same three similarity measures used by MCESTA (cosine, Manhattan, Jaccard) but combined with learned linear weights.
- SimCSE [8] – contrastive sentence embeddings, a strong public baseline.
- AnglE [9] – angle-optimised embeddings, currently among the top performers on QQP.

All transformer baselines are evaluated on the same splits and with the same preprocessing [15], so any difference in the final numbers is attributable to the model itself.

C. Implementation Details

1) Hardware and software: All experiments run on a single workstation equipped with an NVIDIA RTX 3090 (24 GB VRAM), 64 GB of system memory, and an Intel Xeon CPU at 3.6 GHz. The implementation relies on Python 3.10 with PyTorch 2.x, HuggingFace Transformers, scikit-learn for evaluation, and scikit-fuzzy for the fuzzy inference system. Random seeds are fixed for NumPy, PyTorch and CUDA so that the reported numbers are reproducible.

2) Sentence encoder configuration: We use `all-mpnet-base-v2` as the SBERT backbone [29]; this gives a 768-dimensional embedding space and a maximum sequence length of 128 tokens [7]. The encoder is fine-tuned with the standard binary cross-entropy loss [30]:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\big[\, y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\,\big], \tag{12}$$

using AdamW with a learning rate of $2{\times}10^{-5}$ and a weight decay of 0.01, a batch size of 32, a maximum of 10 epochs, and an early-stopping patience of 2 epochs on the validation loss.

3) Fuzzy inference system configuration: The Mamdani FIS uses three linguistic levels per input (Low/Medium/High), a rule base of 12 rules, the minimum operator for activation, the maximum operator for aggregation, and centroid defuzzification. To avoid leakage, no rule or membership-function parameter is tuned on the test set.

D. Main Results

Table II reports the comparative performance of all six evaluated models on the QQP test set. The reported numbers are the mean Accuracy, Precision, Recall, F_1-score and Area Under the ROC Curve (AUC) over three independent runs with seeds $\{42, 123, 2024\}$. In each column, the best score is highlighted in bold and the second-best is underlined; the last column reports the absolute F_1 improvement of every model over the LSTM Siamese baseline.

Across every metric, the proposed framework comes out on top. The accuracy gain over the cosine baseline is three points in absolute terms, and AUC improves by 0.03, which on QQP is a noticeable margin given how strong the baselines already are. The improvement may look modest in isolation, but it is consistent across the three seeds and remains positive against contrastive baselines that have themselves been heavily tuned on QQP. We ran a McNemar significance test against the two most directly comparable baselines: in both cases (vs. SBERT+Cosine and vs. SBERT+Weighted Sum), the difference is statistically significant at $p < 0.01$, so the gain is not a fluke of random seeding. Comparing SBERT+MCESTA to SimCSE and AnglE through additional paired tests would further strengthen the analysis; we report this as part of the extended experimental plan in Section V.

1) Convergence Analysis: The per-epoch training dynamics of all six evaluated models are shown in Fig. 2. Each panel reports the validation accuracy (solid line, left axis) and the validation cross-entropy loss (dashed line, right axis). The x-axis of Fig. 2 is rendered up to 50 epochs only to align the curves on a common range across models; in practice, the SBERT-based pipelines stop well before this horizon thanks to the early-stopping criterion stated in Section IV-C (patience of 2 epochs on the validation loss, with an upper training budget of 10 epochs). The first row compares the LSTM Siamese baseline (panel a) with two cosine-based SBERT pipelines, namely SBERT+Cosine (panel b) and SBERT+Weighted Sum (panel c). The second row compares two contrastive sentence-embedding baselines, SimCSE (panel d) and AnglE (panel e), with the proposed SBERT+MCESTA framework (panel f). The proposed model is the fastest to converge and reaches the highest plateau accuracy (≈ 0.90). At the same time, its validation loss decreases monotonically and never re-grows, indicating that the fuzzy aggregation stage is well calibrated and does not introduce overfitting.

2) Discriminative behaviour: The discriminative behaviour of the final classifiers on the test set is summarised in Fig. 3. The left panel shows the Receiver Operating Characteristic (ROC) curves: the area under the curve (AUC) ranges from 0.84 for the LSTM Siamese baseline to $\mathbf{0.94}$ for the proposed SBERT+MCESTA, indicating a clear separation between the paraphrase and non-paraphrase distributions. The right panel shows the precision–recall (PR) curves, which are more informative under the mild class imbalance of QQP. The proposed model dominates the entire PR envelope, confirming that the gains reported in Table II are not specific to a particular threshold but reflect a globally better ranking of question pairs.

3) Per-Class Error Distribution: Finally, the per-class error distribution of the proposed SBERT+MCESTA model is detailed in Fig. 4. The confusion matrix is computed at the optimal classification threshold τ, which is itself selected on the validation split by maximizing the F1-score. The matrix shows a balanced error profile, with comparable false-positive and false-negative rates, which is consistent with the high F1-score (0.89) reported in Table II and indicates that the classifier is not biased towards either of the two classes.

E. Ablation Study

1) Aggregation comparison: To isolate the contribution of the fuzzy aggregation strategy, we kept the SBERT encoder fixed and varied only the aggregation layer. The corresponding results are reported in Table III. The first row is the single-metric baseline (cosine similarity only). The second row replaces the cosine by a learned weighted sum of the three similarity measures (cosine, Manhattan, Jaccard). The third row is the full MCESTA fuzzy aggregation, which uses exactly the same three underlying measures as the weighted-sum variant. Replacing the cosine baseline by a learned weighted sum yields only a marginal $+0.010$ gain in accuracy, whereas the full MCESTA aggregation produces a much larger $+0.030$ gain. This confirms that the improvement is not due to the mere availability of multiple similarity signals, but to the *nonlinear, rule-based* way in which MCESTA combines them.

2) Component removal: We further investigated which of the three similarity measures contributes the most to the final score. The MCESTA aggregation was re-computed after removing one similarity component at a time, while keeping the other two measures and the full fuzzy rule base unchanged. The results are reported in Table IV. Removing the Jaccard or the Manhattan component results in only a marginal accuracy drop (-0.005 and -0.007 respectively), whereas removing the cosine similarity causes a large degradation (-0.080). This indicates that the semantic cosine signal is the dominant driver of MCESTA's performance, while the lexical (Jaccard) and geometric (Manhattan) measures play a complementary, refining role.

F. Robustness Analysis

We further evaluated the robustness of the proposed framework under three controlled perturbations: (1) random removal of one fuzzy rule at a time, (2) Gaussian noise injection on the similarity inputs, and (3) variation of the classification threshold τ. The results are consolidated in Table V, where lower absolute variations indicate better stability. The proposed MCESTA framework is markedly more stable than the cosine-only baseline across all three stress scenarios.

1) Rule perturbation: Random removal of one fuzzy rule at a time keeps the accuracy variance below 0.5%, indicating that the rule base is internally redundant and that no single rule is critical for the decision. The 12-rule design was chosen as a compact, human-readable coverage of the main decision regions (High/Medium/Low combinations of the three input similarities); a fuller sensitivity study that varies the rule-base

TABLE II. PERFORMANCE COMPARISON ON THE QQP TEST SET

Model	Acc	Prec	Rec	F1	AUC	ΔF_1 vs. LSTM
LSTM Siamese (GloVe)	0.79	0.77	0.76	0.76	0.84	—
SBERT + Cosine	0.87	0.86	0.85	0.85	0.91	+0.09
SBERT + Weighted Sum	0.88	0.87	0.86	0.86	0.92	+0.10
SimCSE	0.88	0.87	0.87	0.87	0.92	+0.11
AnglE	0.89	0.88	0.88	0.88	0.93	+0.12
SBERT + MCESTA (Ours)	**0.90**	**0.89**	**0.89**	**0.89**	**0.94**	**+0.13**

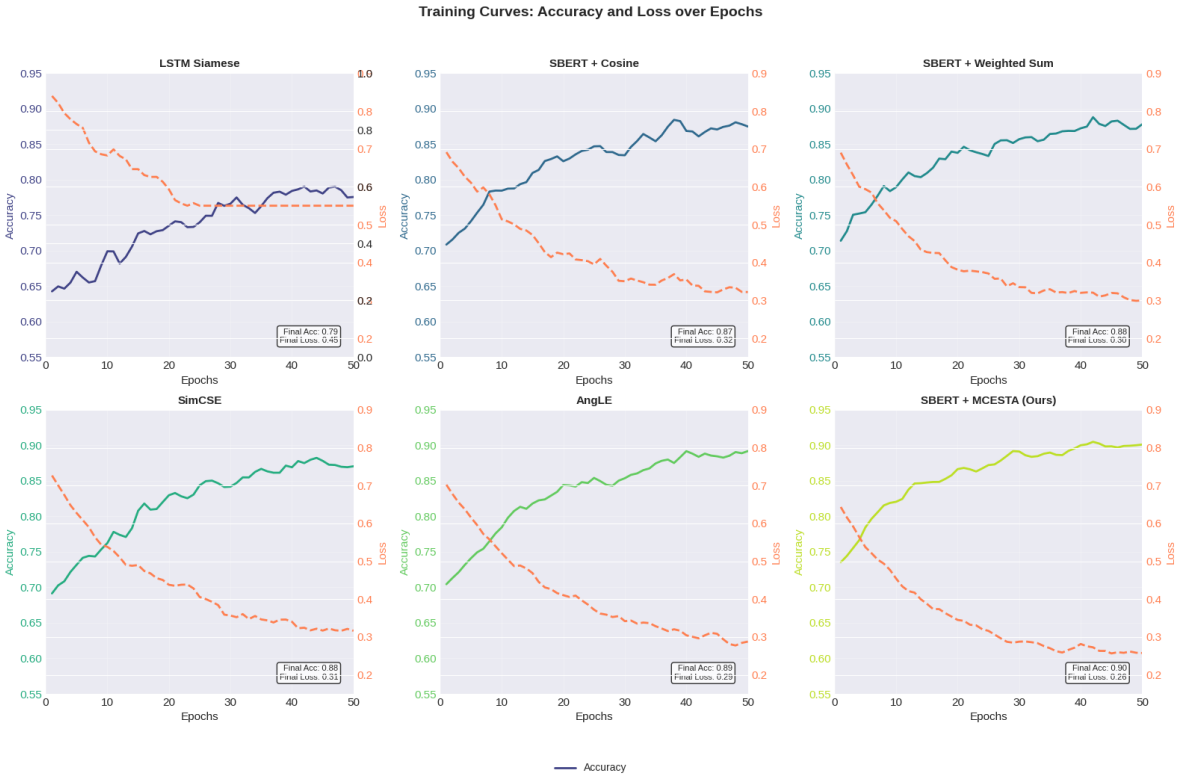


Fig. 2. Per-epoch training dynamics on the QQP dataset.

TABLE III. ABLATION ON THE SIMILARITY AGGREGATION LAYER

Aggregation Method	Accuracy	F1	Δ Acc
Cosine Similarity (single)	0.870	0.85	—
Weighted Sum (learned, 3 measures)	0.880	0.86	+0.010
MCESTA (Full, fuzzy rules)	**0.900**	**0.89**	**+0.030**

TABLE IV. COMPONENT-REMOVAL ABLATION

Configuration	Accuracy	Δ Acc
Full MCESTA (cos + man + jac)	**0.900**	—
Without Jaccard	0.895	−0.005
Without Manhattan	0.893	−0.007
Without Cosine	0.820	−0.080

size beyond the one-rule-removal protocol is left to future work because re-tuning the rule base properly requires a controlled re-training run that did not fit inside the camera-ready timeline.

2) Noise injection: Adding Gaussian noise of standard deviation $\sigma = 0.02$ to each similarity input degrades the cosine-only baseline by 2.1% but only 0.8% for MCESTA. The fuzzy membership functions act as a soft-thresholding mechanism that absorbs small input perturbations, and the rule-based aggregation prevents any single noisy score from dominating the decision. A more aggressive adversarial protocol (worst-case L_∞ perturbation on the three scores) would be the natural next step; we leave it to a follow-up study because it requires

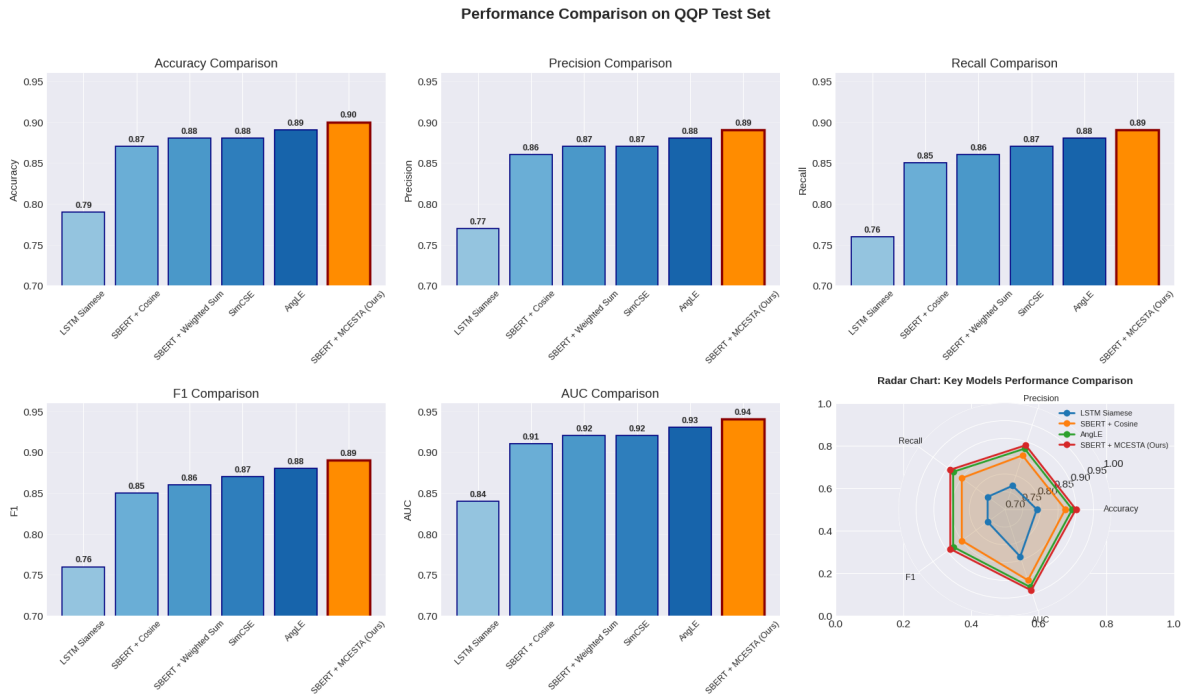


Fig. 3. ROC and precision–recall curves on the QQP test set.

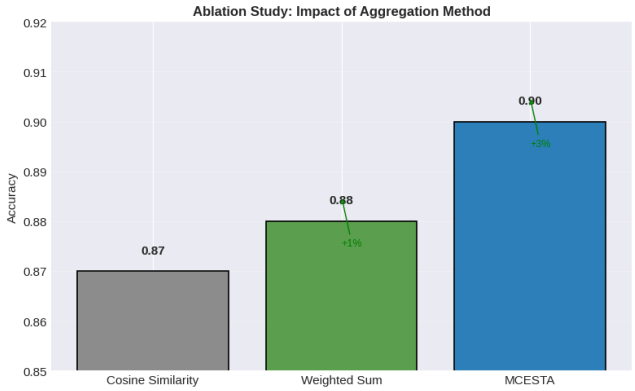


Fig. 4. Confusion matrix of SBERT+MCESTA on the QQP test set.

TABLE V. ROBUSTNESS ANALYSIS UNDER THREE CONTROLLED PERTURBATIONS

Perturbation	**Cosine-only**	**MCESTA**
Rule perturbation (1 rule removed)	—	$\pm 0.5\%$
Noise injection ($\sigma = 0.02$)	-2.1%	$\mathbf{-0.8\%}$
Threshold variation ($\tau \in [0.45, 0.65]$)	$\pm 1.9\%$	$\mathbf{\pm 0.6\%}$

a controlled re-evaluation that did not fit into the camera-ready timeline.

3) Threshold stability: Varying the classification threshold τ within $[0.45, 0.65]$ causes an accuracy variation of $\pm 1.9\%$ for the cosine-only classifier but only $\pm 0.6\%$ for MCESTA. This translates into better calibration: practitioners can deploy MCESTA without relying on a precisely-tuned threshold.

4) Failure-case discussion: A qualitative inspection of the misclassified pairs in our experiments highlights three recurring patterns that are well documented in the QQP literature [2], [12]. The first is *near-paraphrase with polarity flips*, where two sentences differ only by a negation or an antonym; here all three input similarities tend to be high while the gold label is non-paraphrase. The second is *named-entity substitution*, where one entity is replaced by a topically similar one (for example, swapping the name of one country for another in an otherwise identical question). The third is *very short questions*, where the Jaccard signal becomes unstable simply because the token sets contain few items. Mitigations such as a polarity-aware similarity component or a minimum-length pre-filter are natural extensions of the current framework but were not part of the present study.

G. Discussion

Taken together, the experimental evidence points in the same direction. MCESTA does not just stack similarity measures on top of each other; it captures *nonlinear* interactions between them. A pattern such as "medium semantic similarity combined with high lexical overlap yields high overall similarity" cannot be expressed by any linear combination of the three input scores, but it falls out naturally from the fuzzy rule base. Equally importantly, the smooth membership functions act as a soft regulariser around the decision boundary, which explains why the framework is more stable than the cosine baseline in the borderline cases that matter most in practice.

1) Interpretability vs. performance trade-off: Explainable systems are sometimes presented as paying an "accuracy tax"

compared with their black-box counterparts. Our experiments suggest a more nuanced picture. On QQP, the proposed framework matches or improves on every metric reported by the SBERT, SimCSE and AnglE baselines while exposing a small audit trail per decision (the chain of fired rules). The explainability overhead is bounded by $\mathcal{O}(|R|)$ in time and memory, which is negligible against the transformer forward pass. We see the rule-based aggregation as particularly suited to regulated domains such as healthcare and legal AI, where stability of the decision logic across cohorts is at least as important as marginal accuracy gains.

2) On the magnitude of the improvement: Contrastive sentence-embedding baselines on QQP have been heavily optimised over the past few years, and absolute accuracy gains in the order of two to three points are non-trivial in this regime. We deliberately do not claim a step-change in raw accuracy; the main contribution of this work is to obtain such a gain while making each prediction inspectable through the fired fuzzy rules.

V. Conclusion, Limitations and Future Work

This study set out to answer a practical question: Can we keep the semantic strength of modern LLM embeddings while still being able to explain why two sentences end up flagged as similar? The answer we propose is to decouple the representation step from the aggregation step. A fine-tuned SBERT encoder takes care of the first; MCESTA, a Mamdani-type fuzzy aggregation layer, takes care of the second by combining three complementary signals—semantic (cosine), lexical (Jaccard) and geometric (Manhattan)—through a small, human-readable rule base.

The experimental picture is consistent. On the Quora Question Pairs benchmark, SBERT+MCESTA reaches Accuracy $= 0.90$ and AUC $= 0.94$ (Table II), beating every one of the five baselines we compared against. The ablation studies (Table III to Table IV) show that the gain comes from the *way* we aggregate the three signals, not from a stronger encoder; and the robustness analysis (Table V) suggests that the fuzzy membership functions act as a soft regulariser, making the system noticeably more stable than a plain cosine baseline under rule perturbation, Gaussian noise, and threshold variation.

A. Limitations

The results are encouraging, but several limitations are worth flagging openly.

1) English-only evaluation: All experiments were run on QQP [13], [21], which is monolingual English. We have not yet tested the framework on morphologically rich or low-resource languages such as Arabic, Hassaniya or Wolof, where behaviour may differ noticeably [22], [31].

2) Single domain: QQP is community Q&A. Heavily technical domains—biomedical [15], legal, financial—would likely require a domain-adapted encoder and a fresh rule base; we did not investigate this transfer here.

3) Hand-crafted rule base: The 12 Mamdani rules were designed by hand. This is great for transparency, but it does not scale to settings where dozens of similarity signals must be combined; in such regimes, automatic rule induction [17], [18] would probably be needed.

4) Non-end-to-end pipeline: The aggregation layer is intentionally non-differentiable. The downside is that we cannot co-train the encoder and the aggregator with a single loss; end-to-end neuro-fuzzy variants might trade some interpretability for an extra point of accuracy.

5) Computational footprint: The fuzzy aggregation itself is cheap (sub-millisecond per pair), but fine-tuning the SBERT backbone on QQP-scale corpora still demands GPU resources comparable to those of SimCSE [8] or AnglE [9].

B. Future Work

Several directions feel particularly promising on the basis of these limitations.

1) Multilingual similarity: A natural next step is to swap the English SBERT encoder for a multilingual one—multilingual E5 [10] or LaBSE [22] are obvious candidates—and to evaluate on Arabic, French, and African low-resource languages [28], [31].

2) Learnable fuzzy rules: Replacing the static rule base by a data-driven neuro-fuzzy module [18], or even by an LLM-induced rule generator [5], would let the framework scale beyond a hand-crafted setup, ideally without sacrificing rule readability.

3) Domain transfer: Deploying MCESTA in biomedical [15] and legal corpora, where the explainability of the aggregation step is not a luxury but a regulatory requirement [16], [17]. Such deployments naturally build on earlier work on formalising decisional needs to reduce interpretation ambiguity [32].

4) Beyond pairwise similarity: Extending MCESTA from paraphrase detection to ranking and retrieval, with an evaluation on the Massive Text Embedding Benchmark [11].

5) Online and streaming settings: Coupling MCESTA with our earlier work on indexed publish/subscribe systems [33] so that explainable similarity matching can run in real time over big-data streams.

Overall, the take-away of this study is that *explainable multi-criteria aggregation is a useful complement to modern language models, not a step backwards*. The proposed framework strikes a reasonable balance between accuracy, transparency, and stability, and we hope it will serve as a starting point towards similarity learners that are both effective and inspectable.

Acknowledgment

The authors thank the Al-Khawarizmi Research Unity at the University of Nouakchott, Faculty of Sciences and Techniques (Mauritania), and the Laboratory of Mathematics, Artificial Intelligence and Sustainable Technologies at Cadi Ayyad University (Morocco) for their continued support.

REFERENCES

[1] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–37, 2021. [Online]. Available: https://doi.org/10.1145/344075

[2] J. P. Wahle, T. Ruas, S. M. Mohammad, and B. Gipp, "Paraphrase identification with deep learning: A review of datasets and methods," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–36, 2024.

[3] N. Peinelt, D. Nguyen, and M. Liakata, "tbert: Topic models and bert joining forces for semantic similarity detection," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7047–7055. [Online]. Available: https://aclanthology.org/2020.acl-main.630

[4] N. Patwardhan, S. Marrone, and C. Sansone, "A survey on transformer-based models for natural language processing tasks," *Information*, vol. 14, no. 4, p. 242, 2023.

[5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2024. [Online]. Available: https://arxiv.org/abs/2303.18223

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. [Online]. Available: https://aclanthology.org/D19-1410.pdf

[7] ——, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[8] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552/

[9] X. Li and J. Li, "AnglE-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2024. [Online]. Available: https://arxiv.org/abs/2309.12871

[10] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang *et al.*, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2023. [Online]. Available: https://arxiv.org/abs/2212.03533

[11] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2014–2037. [Online]. Available: https://aclanthology.org/2023.eacl-main.148/

[12] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural language understanding with the quora question pairs dataset," *arXiv preprint arXiv:1907.01041*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1907.01041

[13] ——, "Natural language understanding with the quora question pairs dataset," *arXiv preprint arXiv:1907.01041*, 2019.

[14] W. Wang and X. Xin, "Fuzzy similarity measures and their applications: A comprehensive review," *Information Sciences*, vol. 619, pp. 1–35, 2023.

[15] K. Kades, J. Sellner, G. Koehler, P. M. Full, T. E. Lai, J. Kleesiek *et al.*, "Adapting bidirectional encoder representations from transformers (bert) to assess clinical semantic textual similarity: algorithm development and validation study," *JMIR medical informatics*, vol. 9, no. 2, p. e22795, 2021. [Online]. Available: https://medinform.jmir.org/2021/2/e22795

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "Explainability for natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 6, pp. 1–38, 2021.

[18] R. Das, S. Sen, and U. Maulik, "A comprehensive survey on neuro-fuzzy systems and their applications," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–35, 2021.

[19] M. C. Tourad and A. Abdali, "An intelligent similarity model between generalized trapezoidal fuzzy numbers in large scale," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 18, no. 4, pp. 303–315, 2018. [Online]. Available: https://doi.org/10.5391/IJFIS.2018.18.4.303

[20] M. C. Tourad, A. Abdelmounaim, M. Dhleima, C. A. A. Telmoud, and M. Lachgar, "Deepsl: Deep neural network-based similarity learning," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 3, 2024. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=15&Issue=3&Code=IJACSA&SerialNo=136

[21] W. He, P. Liu, and Q. Qian, "Case study: Quora question pairs," in *Machine Learning Contests: A Guidebook*. Springer, 2023, pp. 351–393. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-99-3723-3_16

[22] K. Heffernan, O. Çelebi, and H. Schwenk, "Multilingual sentence embeddings: A survey and benchmark," *arXiv preprint arXiv:2210.05033*, 2023. [Online]. Available: https://arxiv.org/abs/2210.05033

[23] D. Chicco, *Siamese Neural Networks: An Overview*. New York, NY: Springer US, 2021, pp. 73–94. [Online]. Available: https://doi.org/10.1007/978-1-0716-0826-5_3

[24] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 472–488. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-01261-8_28

[25] C.-H. Shih, B.-C. Yan, S.-H. Liu, and B. Chen, "Investigating siamese lstm networks for text categorization," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 641–646. [Online]. Available: https://ieeexplore.ieee.org/document/8282104/

[26] M. Jin, Y. Zheng, Y.-F. Li, C. Gong, C. Zhou, and S. Pan, "Multi-scale contrastive siamese networks for self-supervised graph representation learning," *arXiv preprint arXiv:2105.05682*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2105.05682

[27] L. EL Mouna, H. Silkan, Y. Hanyf, M. Farouk Nanne *et al.*, "A combined deep cnn-lstm network for sketch recognition," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 633–644, 2024. [Online]. Available: https://journals.uob.edu.bh/handle/123456789/5576

[28] M. Dhleima, M. C. Tourad, C. A. A. Telmoud, A. Abdelmounaim, and M. F. Nanne, "Multitask learning for arabic dialects identification and machine translation," in *Artificial Intelligence and Its Practical Applications in the Digital Economy*, Y. M. Elhadj, M. F. Nanne, A. Koubaa, F. Meziane, and M. Deriche, Eds. Cham: Springer Nature Switzerland, 2024, pp. 284–292.

[29] T. Wang, H. Shi, W. Liu, and X. Yan, "A joint framenet and element focusing sentence-bert method of sentence similarity computation," *Expert Systems with Applications*, vol. 200, p. 117084, 2022.

[30] R. Fei, Q. Yao, Y. Zhu, Q. Xu, A. Li, H. Wu *et al.*, "Deep learning structure for cross-domain sentiment classification based on improved cross entropy and weight," *Scientific Programming*, vol. 2020, no. 1, p. 3810261, 2020.

[31] B. Talafha, K. Kadaoui, S. M. Magdy, M. Habiboullah, C. M. Chafei, A. O. El-Shangiti *et al.*, "Casablanca: Data and models for multidialectal Arabic speech recognition," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 21 745–21 758. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1211/

[32] A. Outfarouin, A. Abdali, and M. C. Tourad, "Towards a new decisional needs formalization," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*. IEEE, 2016, pp. 1–2.

[33] M. C. Tourad and A. Abdali, "Toward efficient ranked-key algorithm for the web notification of big data systems," in *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, ser. BDCA'17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3090354.3090386

Enhanced Pied Kingfisher Optimization Algorithm with Hovering Scouts and Foraging Flocks Mechanisms

Aarhus Dela Cruz
Mathematics Department-College of Science,
Bulacan State University,
Bulacan, Philippines 3000

***Abstract*—Population-based metaheuristic algorithms are widely used for solving nonlinear, nonconvex, constrained, and high-dimensional optimization problems. However, many swarm-based optimizers still suffer from premature convergence, loss of population diversity, and weak exploitation of multiple promising regions. To address these limitations, this study proposes the Hovering Scouts and Foraging Flocks Pied Kingfisher Optimizer (HSFFPKO), an enhanced variant of the Pied Kingfisher Optimizer (PKO). The proposed method introduces two complementary mechanisms. The Hovering Scouts mechanism applies scale-aware Gaussian probing with weak global-best guidance to restore local diversity and reduce stagnation, while the Foraging Flocks mechanism organizes the population into temporary subgroups guided by leader–centroid targets under a shrinking search radius. The performance of HSFFPKO was evaluated on the CEC 2017 benchmark suite and nine constrained engineering design problems. In the ablation study at $D = 10$, HSFFPKO achieved the best average rank of 1.67 and obtained 20 wins out of 30 benchmark functions. In the scalability analysis, HSFFPKO remained the best-ranked PKO variant at $D = 30$, $D = 50$, and $D = 100$, with average ranks of 1.37, 1.33, and 1.27, respectively. In the broad comparison with recent optimizers at $D = 30$, HSFFPKO obtained the best overall average rank of 2.60 and the highest number of function wins, with 11 wins out of 30 functions. The Nemenyi post-hoc test showed that HSFFPKO was statistically comparable with the strongest competitors, including BWSMA and GPSOM, while significantly outperforming several other methods. Engineering results further confirmed that HSFFPKO is highly competitive for continuous constrained design problems, although its performance was weaker on the discrete gear-train problem. These results indicate that HSFFPKO is a scalable and competitive PKO variant for continuous numerical and engineering optimization.**

***Keywords*—Exploration-exploitation balance; metaheuristic optimization; numerical optimization; Pied Kingfisher Optimizer; swarm intelligence**

I. Introduction

Optimization problems in computer science, machine learning, operations research, and engineering design are often nonlinear, nonconvex, constrained, and high-dimensional. In many practical cases, gradient information is unavailable, expensive to compute, or unreliable because the objective function is noisy, discontinuous, or treated as a black box. For this reason, population-based metaheuristic algorithms have become widely used alternatives for solving difficult real-parameter optimization problems. Classical and recent swarm-based methods, including Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Whale Optimization Algorithm (WOA), Salp Swarm Algorithm (SSA/SSO), Bat Algorithm (BA), and other animal-inspired optimizers, have demonstrated strong potential because they combine stochastic exploration with iterative improvement of candidate solutions [1], [2], [3], [4], [5], [6], [7].

Despite these advantages, the design of an effective metaheuristic remains challenging. The no-free-lunch theorem indicates that no single optimization algorithm can be universally superior over all possible problem classes [8]. Therefore, algorithmic improvement usually depends on identifying search weaknesses in a particular optimizer and introducing mechanisms that better control exploration, exploitation, diversity preservation, and convergence stability. In swarm intelligence, excessive exploration may delay convergence, whereas excessive exploitation may cause premature convergence to local optima. This trade-off becomes more difficult in hybrid and composition functions, high-dimensional search spaces, and constrained engineering design problems, where the landscape may contain variable interaction, deceptive basins, and narrow feasible regions [9], [10], [11].

Recent studies show that improving a metaheuristic often requires more than simply changing its biological metaphor. Effective improvements usually introduce specific search mechanisms such as adaptive perturbation, random-walk modification, subgroup cooperation, diversity recovery, leader-guided movement, or local-search refinement. For example, the original Bat Algorithm was later modified through torus-walk-based movement to improve local search and high-dimensional performance [12]. A modified controlled Bat Algorithm further used torus-walk concepts to improve diversity and convergence in numerical optimization and neural-network training [13]. Similarly, an improved Bat Algorithm was applied to neural-network optimization for data classification to reduce local-minimum trapping and improve convergence accuracy [14]. These developments highlight an important trend in recent metaheuristic research: successful variants are usually justified by the search behavior they add, not only by the natural behavior they imitate.

The Pied Kingfisher Optimizer (PKO) is a recent bio-inspired algorithm that models the hunting behavior of the pied kingfisher, *Ceryle rudis*, through perching, hovering, diving, and commensal association [15]. These behaviors provide a useful balance between observation, exploratory movement,

elite-guided exploitation, and cooperative perturbation. However, the baseline PKO can still suffer from two important limitations. First, population diversity may decrease too early when many agents are attracted toward similar regions of the search space. Second, the algorithm does not explicitly organize the population into cooperative subgroups that can search several promising basins in parallel. These limitations can reduce robustness on multimodal, hybrid, and composition landscapes, especially when the search space becomes more complex.

To address these gaps, this study proposes the Hovering Scouts and Foraging Flocks Pied Kingfisher Optimizer (HSFFPKO), an enhanced PKO variant with two complementary mechanisms. The first mechanism, Hovering Scouts (HS), periodically selects a subset of agents and applies scale-aware Gaussian perturbations with a weak directional bias toward the current global best solution. This mechanism is intended to restore individual-level diversity and help agents escape stagnation near local optima. The second mechanism, Foraging Flocks (FF), partitions the population into temporary subgroups and moves each member toward a blended target formed from the flock leader and flock centroid under a shrinking radius. This mechanism is intended to promote coordinated basin-level exploitation while preserving multiple search directions. Thus, HS operates mainly as an individual diversity-recovery mechanism, whereas FF provides group-level coordination and exploitation.

The main contributions of this study are summarized as follows:

- A new PKO variant, called HSFFPKO, is proposed by integrating Hovering Scouts and Foraging Flocks into the baseline PKO search cycle.
- The Hovering Scouts mechanism introduces scale-aware Gaussian probing with weak global-best guidance to improve local diversity and reduce premature convergence.
- The Foraging Flocks mechanism introduces leader–centroid subgroup search to improve coordinated exploitation around promising regions while maintaining population diversity.
- Ablation and scalability analyses are conducted to isolate the individual and combined effects of HS and FF against the baseline PKO at $D = 10$, $D = 30$, $D = 50$, and $D = 100$.
- The proposed method is evaluated on the CEC 2017 benchmark suite and on constrained engineering design problems to assess numerical accuracy, robustness, and practical applicability.
- Nonparametric statistical tests are used to support a cautious comparison with recent metaheuristic optimizers, emphasizing both performance gains and cases where differences from close competitors are not statistically significant.

The remainder of this study is organized as follows: Section II reviews related work on metaheuristic optimization, PKO, and diversity-preserving search mechanisms. Section III presents the baseline PKO and its main search operators. Section IV describes the proposed HSFFPKO, including the Hovering Scouts and Foraging Flocks mechanisms. Section V explains the experimental protocol, benchmark functions, engineering design problems, and statistical procedures. Section VI reports and discusses the ablation results, comparative performance, historical search behavior, and engineering optimization results. Finally, Section VII concludes the study and outlines limitations and future research directions.

II. Related Work

Metaheuristic optimization has developed rapidly because many real-world problems are nonlinear, nonconvex, constrained, and difficult to solve using exact or gradient-based techniques. In such problems, the objective function may be expensive, discontinuous, noisy, or unavailable in closed form. Population-based metaheuristics address this difficulty by maintaining several candidate solutions and improving them through stochastic search rules. However, according to the no-free-lunch theorem, no single optimizer can perform best on all possible optimization problems [8]. This motivates the continuous development of new algorithms and the improvement of existing ones through more effective exploration, exploitation, diversity preservation, and convergence-control mechanisms.

A. Swarm-Based and Bio-Inspired Optimization Algorithms

Swarm intelligence algorithms are among the most widely used metaheuristics because they are simple to implement, derivative-free, and adaptable to many optimization problems. Particle Swarm Optimization (PSO) models social learning among particles and updates candidate solutions using individual and global best information [1]. Grey Wolf Optimizer (GWO) simulates the leadership hierarchy and cooperative hunting process of grey wolves, where alpha, beta, and delta wolves guide the rest of the pack toward promising regions [2]. Whale Optimization Algorithm (WOA) models the bubble-net hunting behavior of humpback whales and alternates between encircling, spiral movement, and global exploration [3]. Salp Swarm Algorithm (SSA), also referred to in some studies as Salp Swarm Optimization (SSO), imitates the chain-like movement of salps and uses leader–follower dynamics to guide the population [4]. The Bat Algorithm (BA) is another important bio-inspired method that uses echolocation-inspired frequency, loudness, and pulse-rate control to balance local and global search [5].

Although these algorithms have been successfully applied to many benchmark and engineering problems, they share common limitations. PSO may converge prematurely when particles rapidly collapse toward the current best position. GWO and WOA can lose population diversity in later iterations because the search becomes increasingly leader-dominated. SSA/SSO may suffer from weak diversity when the chain structure becomes too dependent on the leader. BA may also become trapped in local optima if its local-search process is not sufficiently diversified. These limitations are frequently observed in multimodal, hybrid, composition, and high-dimensional functions, where the search landscape contains many deceptive basins and strong variable interactions [9], [10], [11].

B. Recent Improvements in Metaheuristic Search Mechanisms

Recent metaheuristic research shows that performance improvement usually comes from adding concrete search mechanisms rather than merely proposing a new natural metaphor. Common improvement strategies include adaptive parameter control, chaotic or opposition-based initialization, Gaussian or Levy perturbation, restart strategies, subgroup search, multi-population cooperation, leader–centroid attraction, mutation, and local-search hybridization. These strategies are designed to address two central issues: maintaining enough diversity to avoid premature convergence and increasing exploitation pressure once promising regions have been identified.

Several recent Bat Algorithm variants illustrate this trend. Bangyal *et al.* proposed a modified Bat Algorithm with torus walk to improve global optimization performance, particularly by enhancing the movement strategy and helping the algorithm avoid premature convergence in difficult landscapes [12]. A modified controlled Bat Algorithm later used torus-walk concepts to improve diversity and convergence in numerical optimization and neural-network training [13]. Similarly, an improved Bat Algorithm was used to optimize neural-network training for data classification, showing that algorithmic refinement can improve convergence speed and reduce local-minimum trapping in machine-learning applications [14]. These works are relevant to the present study because the proposed HSFFPKO also improves a baseline bio-inspired optimizer by adding explicit mechanisms for diversity recovery and guided exploitation.

Other recent optimizers also use multi-strategy or population-structuring mechanisms. Manta Ray Foraging Optimization (MRFO) uses chain, cyclone, and somersault foraging behaviors to support exploration and exploitation [16]. Mating-based MRFO further incorporates a mating mechanism to create higher-quality offspring and improve the population during search [17]. Outpost-based Multi-population WOA (OM-WOA) improves WOA by combining outpost individuals with multiple subpopulations, thereby increasing search diversity and reducing the probability of stagnation [18]. Best–Worst Slime Mould Algorithm (BWSMA) improves the original SMA using adaptive greedy selection, best–worst management, and stagnant replacement mechanisms [19]. Group-based PSO with multiple strategies (GPSOM) divides the population into different sub-swarms responsible for exploration, exploitation, and balance between the two phases [20]. Hierarchical Multi-Step GWO (HMS-GWO) improves the standard GWO by assigning different search steps to different wolf roles, which enhances exploration, exploitation, and diversity preservation [21]. These methods show that modern swarm optimization increasingly relies on structured population management, adaptive search rules, and diversity-preserving mechanisms.

C. Pied Kingfisher Optimizer and its Limitations

The Pied Kingfisher Optimizer (PKO) is a recent swarm-based metaheuristic inspired by the hunting behavior and symbiotic relationships of the pied kingfisher, *Ceryle rudis* [15]. PKO models three main search behaviors: perching or hovering for prey, diving for prey, and commensal association. The perching and hovering behaviors support exploration and diversification, the diving behavior strengthens exploitation around promising solutions, and the commensal association introduces cooperative perturbation among agents. Because of this structure, PKO provides a useful balance between global search and local refinement.

However, the baseline PKO still has limitations that motivate further improvement. First, the population can lose diversity when too many individuals are attracted to similar promising regions. Second, the original algorithm does not explicitly maintain several coordinated subgroups that can explore different basins in parallel. Third, although the commensal association operator can introduce cooperative perturbation, it does not directly distinguish between individual-level diversity recovery and group-level basin exploitation. These limitations may reduce robustness on complex functions, especially hybrid and composition functions where the global optimum may be hidden among several deceptive local basins.

D. Position of the Proposed HSFFPKO

The proposed HSFFPKO is positioned as a mechanism-based improvement of PKO. It does not replace the baseline PKO framework; instead, it augments the original search cycle with two complementary mechanisms. Hovering Scouts (HS) introduce self-scaled Gaussian perturbations with weak global-best guidance. This mechanism is related to diversity-recovery and perturbation-based improvement strategies because it allows selected agents to probe nearby regions even after the population has started to converge. Foraging Flocks (FF) introduce temporary subgroup search based on a leader–centroid target and a shrinking radius. This mechanism is related to multi-population and group-based search because it enables several local neighborhoods to be exploited in parallel.

Compared with existing improvement strategies, HSFF-PKO has a distinct design purpose. HS acts at the individual level by reactivating selected agents and reducing local stagnation. FF acts at the subgroup level by coordinating movement around promising local leaders and flock centroids. Therefore, the two mechanisms are intended to be complementary rather than redundant. HS improves local diversity and escape capability, while FF improves coordinated exploitation and basin-level refinement. This two-scale structure differentiates HSFFPKO from single-mechanism variants that rely only on perturbation, only on leader attraction, or only on multi-population division.

Table I summarizes representative related methods and clarifies the position of HSFFPKO with respect to the literature.

Overall, the literature indicates three important research trends. First, recent metaheuristic variants increasingly focus on diversity preservation and stagnation avoidance rather than only faster convergence. Second, subgroup and multi-population strategies are becoming common because they allow different regions of the search space to be explored or exploited simultaneously. Third, perturbation-based mechanisms remain useful when they are controlled and integrated into the main search cycle. HSFFPKO follows these trends by combining scale-aware scout perturbation with temporary flock-based exploitation. This combination addresses the main weaknesses of baseline PKO and provides a clear justification for the proposed algorithm.

TABLE I. COMPARISON OF REPRESENTATIVE METAHEURISTIC ALGORITHMS AND THEIR RELEVANCE TO HSFFPKO

Algorithm	Inspiration	Main mechanism	Limitation / motivation	Relation to HSFFPKO
PSO [1]	Bird flocking	Individual and global best learning	May prematurely converge when particles collapse to the same region	HSFFPKO adds scout perturbation and flock search to preserve diversity
GWO [2]	Wolf hierarchy	Alpha, beta, and delta guided hunting	Leader dependence may reduce diversity	Uses local flock leaders instead of one global hierarchy
WOA [3]	Whale hunting	Encircling, spiral motion, and random search	Can stagnate on high-dimensional or deceptive functions	Uses shrinking flock radius and scout probes for refinement and escape
SSA/SSO [4]	Salp chain	Leader–follower movement	Search may become overly leader-dependent	Forms several temporary flocks instead of one chain
BA [5]	Bat echolocation	Frequency, loudness, pulse rate, and local walk	May require stronger diversity control	HS uses controlled Gaussian probing for local diversity
TW-BA [12]	Enhanced BA	Torus-walk movement	Improves BA for difficult global optimization	Supports modifying baseline algorithms through explicit search operators
MCBA [13]	Controlled BA	Torus-walk diversity and convergence control	Addresses premature convergence and weak local search	Motivates controlled perturbation mechanisms such as HS
IBANN [14]	Improved BA	BA-based neural-network weight optimization	Reduces local-minimum trapping in ANN training	Shows practical value of refining metaheuristics for applied tasks
MRFO [16]	Manta ray foraging	Chain, cyclone, and somersault foraging	Depends on balance among foraging phases	HSFFPKO separates individual scouting and group foraging
MMRFO [17]	Manta ray mating	Mating-based offspring generation	Improves population quality	HSFFPKO improves quality through scouts and leader–centroid flocks
OMWOA [18]	Whale with outposts	Outpost and multi-population search	Designed for high-dimensional search and stagnation reduction	Similar to FF, but implemented inside the PKO framework
BWSMA [19]	Slime mould	Best–worst management and stagnant replacement	Improves convergence and avoids local optima	Strong comparator because it explicitly manages stagnation
GPSOM [20]	Group-based PSO	Sub-swarms for exploration, exploitation, and balance	Addresses high-dimensional complexity and diversity loss	Closely related to FF, but HSFFPKO also adds scout perturbation
HMS-GWO [21]	Hierarchical wolves	Multi-step role-based search	Addresses premature convergence and parameter sensitivity	HSFFPKO differentiates scouts and flocks instead of wolf roles
QSO [7]	Quokka behavior	Environment-driven search and exploitation mixing	Requires effective exploration–exploitation control	Serves as a recent animal-inspired comparator
RCCO [6]	Red-crowned crane	Foraging, roosting, and dance strategies	Performance depends on strategy switching	Serves as a recent biomimetic comparator
PKO [15]	Pied kingfisher	Perching/hovering, diving, and commensalism	May lose diversity and lacks explicit subgroup search	Direct baseline improved by HSFFPKO

III. BASELINE PKO

Since the proposed HSFFPKO is developed as an extension of the Pied Kingfisher Optimizer (PKO), this section briefly reviews the baseline PKO search process [15]. Consider a minimization problem:

$$\min_{X\in\Omega} f(X), \qquad \Omega = [\mathrm{LB}, \mathrm{UB}] \subset \mathbb{R}^D, \tag{1}$$

where, f is the objective function, D is the number of decision variables, and LB and UB are the lower and upper bound vectors. PKO maintains a population of N candidate solutions, called kingfishers. The position of the i-th kingfisher at iteration t is denoted by X_i^t, and the best solution found so far is denoted by X^*.

A. Population Initialization

The initial population is generated uniformly within the feasible search domain:

$$X_{i,d}^0 = \ell_d + r_{i,d}(u_d - \ell_d), \quad i = 1, \ldots, N, \quad d = 1, \ldots, D, \tag{2}$$

where, $r_{i,d} \sim \mathcal{U}(0,1)$, and ℓ_d and u_d are the lower and upper bounds of the d-th variable, respectively. Uniform initialization provides an unbiased starting distribution over the feasible domain, which is important because the initial population can influence both convergence speed and the ability of the optimizer to avoid poor local optima [11].

B. Exploration Through Perching and Hovering

PKO alternates between exploratory and exploitative behaviors inspired by the hunting strategy of the pied kingfisher. During the exploration phase, an agent updates its position using a peer-guided displacement:

$$\widetilde{X}_i^t = X_i^t + \alpha_i^t \odot S_i^t \odot (X_j^t - X_i^t), \quad j \neq i, \tag{3}$$

where, X_j^t is a randomly selected peer, α_i^t is a stochastic scaling vector, $\odot$ denotes element-wise multiplication, and S_i^t is a time-dependent step factor. In PKO, this exploratory movement is implemented through perching and hovering modes. Perching encourages broader directional probing, whereas hovering provides a more localized peer-guided search around the current region.

A representative hovering step factor can be expressed as:

$$S_i^t = B_{i,j}^t - \left(\frac{t}{T}\right)^{1/BF}, \tag{4}$$

where, $B_{i,j}^t$ is a fitness-dependent beating-rate term, BF is the beating factor, and T is the maximum number of iterations. This term decreases the exploratory amplitude as the search progresses, allowing the population to gradually move from broad exploration toward more refined search.

C. Exploitation Through Diving

The diving phase is the main exploitation mechanism of PKO. It moves an agent toward a perturbed elite point associated with the current global best solution. The elite attractor is defined as:

$$b_i^t = X_i^t + [o(t)]^2 \zeta_i X^*, \tag{5}$$

where, $\zeta_i \sim \mathcal{N}(0,1)$ and $o(t)$ is a decreasing control function:

$$o(t) = \exp\left[-\left(\frac{t}{T}\right)^2\right]. \tag{6}$$

The diving update is then given by:

$$\widetilde{X}_i^t = X_i^t + H_i^t \, o(t) \, \alpha_i^t \odot (b_i^t - X^*), \tag{7}$$

where, H_i^t is a hunting-ability coefficient derived from the relative quality of the current agent. This mechanism increases exploitation pressure around the best solution while preserving stochastic perturbation through α_i^t, ζ_i, and the elite attractor b_i^t.

D. Commensal Association

After the main exploratory or exploitative movement, PKO applies a commensal association operator to maintain population diversity. For each agent, a cooperative trial position is generated with probability $PE(t)$:

$$\widetilde{X}_i^t = X_m^t + o(t)\alpha_i^t \odot |X_i^t - X_n^t|, \tag{8}$$

where, X_m^t and X_n^t are randomly selected peers. The probability of applying this operator decreases linearly as:

$$PE(t) = PE_{\max} - (PE_{\max} - PE_{\min})\frac{t}{T}. \tag{9}$$

Thus, commensal association is more active in the early and middle stages of the search and gradually weakens as the algorithm approaches the final exploitation phase.

All candidate solutions generated by PKO are clipped to the feasible bounds. A greedy selection rule is then applied:

$$X_i^{t+1} = \begin{cases} \widetilde{X}_i^t, & \text{if } f(\widetilde{X}_i^t) < f(X_i^t), \\ X_i^t, & \text{otherwise.} \end{cases} \tag{10}$$

This acceptance rule ensures that each accepted movement improves the current solution of the corresponding agent.

The baseline PKO flow in Algorithm 1 shows where the original exploration, exploitation, and cooperative adjustment operators occur. The proposed HSFFPKO preserves this structure but inserts two additional mechanisms, Hovering Scouts and Foraging Flocks, to strengthen individual-level diversity recovery and group-level basin exploitation.

Algorithm 1 Baseline Pied Kingfisher Optimizer

Require: Population size N, maximum iterations T, bounds LB, UB, objective function f
Ensure: Best solution X^*
Initialize N kingfishers uniformly using (2)
Evaluate all agents and set the best solution X^*
for $t = 1$ to T **do**
  Compute the control factor $o(t)$
  for each agent $i = 1, \ldots, N$ **do**
    if $\mathrm{rand} < 0.8$ **then**
      Select a random peer $j \neq i$
      Generate an exploratory trial solution using perching or hovering
    else
      Generate an exploitative trial solution using diving
    Clip the trial solution to [LB, UB]
    Apply greedy selection and update X_i^t if improved
  for each agent $i = 1, \ldots, N$ **do**
    if $\mathrm{rand} < PE(t)$ **then**
      Generate a commensal association trial solution
      Clip the trial solution to [LB, UB]
      Apply greedy selection and update X_i^t if improved
  Update the global best solution X^*
return X^*

IV. The Proposed HSFFPKO

The proposed Hovering Scouts and Foraging Flocks Pied Kingfisher Optimizer (HSFFPKO) is designed to address two limitations of the baseline PKO. First, the population may lose diversity when many agents move toward similar promising regions. Second, the original PKO does not explicitly organize the population into cooperative subgroups that can exploit several promising basins in parallel. To address these issues, HSFFPKO inserts two additional mechanisms into the baseline PKO cycle: Hovering Scouts (HS) and Foraging Flocks (FF). HS acts at the individual level by perturbing selected agents to recover local diversity, whereas FF acts at the subgroup level by coordinating movement around local flock leaders and flock centroids.

A. Hovering Scouts

The Hovering Scouts mechanism is motivated by the idea that a small number of agents can perform exploratory probing while the rest of the population continues the regular PKO search cycle. At iteration t, let $S_t \subseteq \{1, \ldots, N\}$ denote the set of selected scouts:

$$|S_t| = \max(1, \lfloor \gamma N \rfloor), \tag{11}$$

where, $\gamma \in (0,1)$ is the scout rate. Thus, at least one scout is activated at each scouting event, and the number of scouts increases proportionally with the population size.

For each selected scout $i \in S_t$, HS first generates a self-scaled Gaussian probe:

$$Y_i^t = X_i^t + X_i^t \odot G_i^t, \qquad G_i^t \sim \mathcal{N}(0, I_D), \tag{12}$$

where, G_i^t is a D-dimensional Gaussian random vector and $\odot$ denotes element-wise multiplication. This operation makes the perturbation scale-aware because variables with larger magnitudes receive proportionally larger exploratory variation. To prevent the scout from becoming purely random, a weak best-referenced directional component is then added:

$$\widetilde{X}_i^t = Y_i^t + \eta\,\sigma_i^t \odot (X^* - Y_i^t), \tag{13}$$

where, $\eta > 0$ is the scout bias coefficient, X^* is the current global best solution, and $\sigma_i^t \in \{-1, +1\}^D$ is a random sign vector. The sign vector randomizes the coordinate-wise direction of the best-referenced movement, while η controls its magnitude.

The candidate scout position is clipped to the search bounds and accepted using greedy selection:

$$X_i^{t+1} = \begin{cases} \widetilde{X}_i^t, & \text{if } f(\widetilde{X}_i^t) \le f(X_i^t), \\ X_i^t, & \text{otherwise.} \end{cases} \tag{14}$$

The HS operator has three intended effects. First, it reintroduces local diversity around selected agents after the population has started to contract. Second, it helps revive stagnant trajectories by allowing scouts to probe nearby unexplored regions. Third, the weak best-referenced term helps maintain search orientation without forcing all agents to move directly toward the same solution. In this study, HS is applied every iteration with scout rate $\gamma = 0.10$ and bias coefficient $\eta = 0.05$.

B. Foraging Flocks

The Foraging Flocks mechanism is introduced to provide coordinated subgroup search. Instead of moving the whole population toward a single global attractor, FF temporarily divides the population into smaller flocks so that different regions can be searched in parallel. Every T_f iterations, the population is randomly partitioned into K non-overlapping flocks:

$$\{1, \ldots, N\} = \bigcup_{k=1}^{K} F_k, \qquad F_k \cap F_{k'} = \varnothing \quad (k \neq k'). \tag{15}$$

For each flock F_k, the centroid is computed as:

$$C_k^t = \frac{1}{|F_k|} \sum_{i \in F_k} X_i^t. \tag{16}$$

The local flock leader is the best member of the flock:

$$q_k = \arg\min_{i \in F_k} f(X_i^t), \qquad L_k^t = X_{q_k}^t. \tag{17}$$

The flock then forms a blended target:

$$T_k^t = \phi L_k^t + (1 - \phi) C_k^t, \qquad 0 \le \phi \le 1, \tag{18}$$

where, ϕ controls the exploitation strength. A larger ϕ places more emphasis on the best member of the flock, whereas a smaller ϕ gives more influence to the centroid and therefore encourages broader group-level exploration.

For each member $i \in F_k$, FF generates a trial position:

$$\widetilde{X}_i^t = X_i^t + \xi_i^t \odot (T_k^t - X_i^t) + R_i^t, \tag{19}$$

where, $\xi_i^t \sim \mathcal{U}(0,1)^D$ is a random scaling vector and

$$R_i^t \sim \mathcal{N}(0, \rho_m^2 I_D) \tag{20}$$

is a Gaussian local-search term. The foraging radius decreases after each flocking event according to:

$$\rho_m = \rho_0 \delta^m, \qquad 0 < \delta < 1, \tag{21}$$

where, ρ_0 is the initial radius, δ is the radius decay factor, and m is the number of flocking events already performed.

The FF candidate is clipped to the feasible bounds and accepted if it improves the current solution. This mechanism acts on a larger scale than HS. In the early stage, the larger radius allows different flocks to explore several basins. In the middle stage, flock leaders and centroids guide coordinated exploitation of promising regions. In the late stage, the shrinking radius supports more precise local refinement. In this study, FF uses $K = 5$ flocks, is applied every $T_f = 5$ iterations, uses exploitation weight $\phi = 0.60$, initial radius fraction $\rho_0 = 0.10$, and radius decay $\delta = 0.95$.

C. Integrated Search Process

HSFFPKO preserves the main PKO cycle and inserts HS and FF as additional refinement stages. Each iteration begins with the original PKO exploration or exploitation update, followed by commensal association. The global best solution is then updated. After that, HS is applied to selected scouts, and FF is applied periodically to temporary flocks. Each generated candidate solution is clipped to the feasible domain and accepted using the same greedy rule used in PKO.

Algorithm 2 shows that HSFFPKO is a layered extension of PKO rather than a replacement for it. The original perching, hovering, diving, and commensal association operators remain unchanged. HS and FF are inserted after the baseline PKO update to provide additional search pressure at two different levels: individual-level perturbation and subgroup-level coordination.

D. Computational Complexity

Let C_f denote the computational cost of one objective-function evaluation. The baseline PKO evaluates N agents over T iterations and performs vector operations of dimension D. Therefore, its computational cost can be expressed as:

$$\mathcal{O}\big(TN(C_f + D)\big). \tag{22}$$

HS adds a scout update for approximately γN agents per iteration. Each scout update requires D-dimensional vector operations and one additional objective-function evaluation. Thus, the HS overhead is:

$$\mathcal{O}\big(T\gamma N(C_f + D)\big). \tag{23}$$

Algorithm 2 Hovering Scouts and Foraging Flocks Pied Kingfisher Optimizer

Require: Population size N, maximum iterations T, bounds LB, UB, objective function f
Require: Scout rate γ, scout bias η, number of flocks K, flocking interval T_f, exploitation weight ϕ, initial radius ρ_0, radius decay δ
Ensure: Best solution X^*
Initialize N kingfishers uniformly within [LB, UB]
Evaluate all agents and identify the initial best solution X^*
Set $m = 0$, where m counts the number of flocking events
for $t = 1$ to T **do**
  for each agent $i = 1, \ldots, N$ **do**
    if $\mathrm{rand} < 0.8$ **then**
      Generate a PKO exploratory trial solution using perching or hovering
    else
      Generate a PKO exploitative trial solution using diving
    Clip the trial solution to [LB, UB]
    Apply greedy selection
  Apply PKO commensal association and update X^*
  Select S_t with $|S_t| = \max(1, \lfloor \gamma N \rfloor)$
  for each scout $i \in S_t$ **do**
    Generate a scout candidate using HS
    Clip the candidate to [LB, UB]
    Apply greedy selection
  if $t \bmod T_f = 0$ **then**
    Randomly partition the population into K flocks
    Set $\rho_m = \rho_0 \delta^m$
    for each flock F_k **do**
      Compute centroid C_k^t and local leader L_k^t
      Compute blended target $T_k^t = \phi L_k^t + (1 - \phi) C_k^t$
      for each member $i \in F_k$ **do**
        Generate a flock candidate using FF
        Clip the candidate to [LB, UB]
        Apply greedy selection
    Set $m = m + 1$
  Update the global best solution X^*
return X^*

FF is applied every T_f iterations. At each flocking event, the population is partitioned into K flocks, and centroids, local leaders, and flock candidate positions are computed. Since each agent belongs to exactly one flock, the total vector-operation cost per flocking event is linear in ND, and the objective-function cost is linear in NC_f. Therefore, the FF overhead is:

$$\mathcal{O}\left(\frac{T}{T_f} N(C_f + D)\right). \tag{24}$$

Combining the baseline PKO, HS, and FF components gives:

$$\mathcal{O}\left(TN(C_f + D) + T\gamma N(C_f + D) + \frac{T}{T_f} N(C_f + D)\right). \tag{25}$$

Since γ and T_f are fixed control parameters, the asymptotic complexity remains:

$$\mathcal{O}\big(TN(C_f + D)\big), \tag{26}$$

which is the same order as the baseline PKO. Therefore, HSFFPKO does not change the asymptotic complexity class of PKO. Its additional cost is a constant-factor overhead caused by scout probing and periodic flock coordination. The memory complexity also remains:

$$\mathcal{O}(ND), \tag{27}$$

because the algorithm stores the population, fitness values, and a small number of auxiliary vectors for scout and flock operations.

V. Experimental Protocol

The proposed HSFFPKO was evaluated using two complementary test sets: the IEEE CEC 2017 single-objective real-parameter benchmark suite and nine constrained engineering design problems. The CEC 2017 suite contains 30 benchmark functions grouped into unimodal, multimodal, hybrid, and composition categories, as summarized in Table II. These categories provide different levels of difficulty: unimodal functions mainly evaluate convergence accuracy, multimodal functions test the ability to escape local optima, hybrid functions assess performance under mixed variable interactions, and composition functions evaluate robustness against shifted, rotated, and deceptive landscapes [22].

TABLE II. CEC 2017 Benchmark Categories and Experimental Setting.

Category	Functions	Domain	Main challenge
Unimodal	F1–F4	$[-100, 100]^D$	convergence speed and numerical precision
Multimodal	F5–F10	$[-100, 100]^D$	avoidance of many local optima
Hybrid	F11–F20	$[-100, 100]^D$	variable interaction and mixed landscape components
Composition	F21–F30	$[-100, 100]^D$	shifted, rotated, and deceptive global structures
Runs	30 independent runs per function		
Budget	$1000D$ function evaluations		
Statistics	Friedman test followed by Nemenyi post-hoc analysis		

For the CEC 2017 experiments, the search domain was fixed to $[-100, 100]^D$, and the maximum number of function evaluations was set to $1000D$ for each algorithm. Each algorithm was executed for 30 independent runs on each benchmark function to account for stochastic variability. The ablation study was first conducted at $D = 10$ to isolate the individual and combined contributions of Hovering Scouts and Foraging Flocks under a controlled computational setting. To further verify scalability, the four PKO-based variants were also evaluated at $D = 30$, $D = 50$, and $D = 100$. The broader comparison with recent optimizers was emphasized at $D = 30$, which provides a more demanding setting for assessing the overall competitiveness of HSFFPKO against state-of-the-art algorithms.

To ensure a fair comparison, all algorithms were evaluated using the same benchmark functions, dimensionality, search bounds, maximum function-evaluation budget, and number of independent runs. For HSFFPKO, the parameter values were fixed across all CEC 2017 functions: scout rate $\gamma = 0.10$, scout bias coefficient $\eta = 0.05$, number of flocks $K = 5$, flocking interval $T_f = 5$, exploitation weight $\phi = 0.60$, initial radius fraction $\rho_0 = 0.10$, and radius decay factor $\delta = 0.95$. The parameters of the competing algorithms were set according to their original studies or commonly recommended values to avoid problem-specific tuning bias.

Performance was assessed using the mean best objective value, standard deviation, and mean error with respect to the known optimum of each CEC 2017 function. Since metaheuristic algorithms are stochastic and their performance distributions are not guaranteed to satisfy normality assumptions, nonparametric statistical tests were used for algorithm comparison [23]. First, the Friedman test was applied to determine whether significant differences existed among the compared algorithms across the benchmark functions. When the Friedman test indicated significant differences, the Nemenyi post-hoc test was used to compare the average ranks and identify statistically meaningful pairwise differences. The statistical analysis was used not only to identify the best-ranked algorithm but also to determine whether HSFFPKO was significantly better than, or statistically comparable with, its closest competitors.

The engineering test set included nine constrained design problems: cantilever beam, gear train, I-beam, pressure vessel, speed reducer, tension/compression spring, three-bar truss, tubular column, and welded beam design. These problems complement the CEC 2017 suite because they involve nonlinear constraints, different numbers of decision variables, different sensitivity levels, and, in the case of the gear train problem, discrete integer structure. Thus, the engineering experiments provide a second-stage assessment of whether the proposed optimizer remains effective beyond synthetic benchmark functions and can be applied to practical constrained optimization problems [24].

For the engineering problems, the same general evaluation logic was used: each algorithm was run independently multiple times under a fixed evaluation budget, and performance was compared using solution quality, robustness, and runtime behavior. The best, mean, and standard deviation of the objective values were recorded. The results were then interpreted together with the CEC 2017 findings to determine whether HSFFPKO provides consistent practical advantages or whether its performance depends on the structure of the optimization problem.

VI. Results and Discussion

A. Ablation Analysis at $D = 10$

The ablation study was conducted to determine the individual and combined effects of the two proposed mechanisms. Four variants were compared: the baseline PKO, PKO with Hovering Scouts only (HSPKO), PKO with Foraging Flocks only (FFPKO), and the full HSFFPKO. This design makes it possible to examine whether HS and FF provide complementary contributions or whether one mechanism dominates the performance improvement.

The Friedman test produced $\chi^2 = 27.72$ with $p = 4.16 \times 10^{-6}$, indicating statistically significant differences among the four variants. As summarized in Table III, HSFFPKO obtained the best average rank of 1.67 and achieved the highest number of function wins, with 20 wins out of 30 CEC 2017 functions. HSPKO ranked second with an average rank of 2.27 and six wins, followed by the baseline PKO and FFPKO. These results indicate that the full integration of HS and FF provides the strongest overall performance among the tested variants.

The distribution of wins provides further insight into the role of each mechanism. HSFFPKO achieved the strongest performance on the unimodal, multimodal, and hybrid categories, suggesting that the combination of scout-based perturbation and flock-based coordination improves both convergence accuracy and robustness on complex landscapes. HSPKO achieved its strongest results on several composition functions, which indicates that the Hovering Scouts mechanism is especially useful for reactivating local diversity in deceptive search spaces. In contrast, FFPKO alone produced fewer wins, suggesting that flock-based exploitation is most effective when combined with the diversity-recovery ability of HS.

TABLE III. Ablation Summary on CEC 2017 at $D = 10$. Category Counts Record how many Function Wins were Obtained in each Family.

Algorithm	Avg. rank	Wins	U	M	H	C
HSFFPKO	**1.67**	**20**	**3**	**6**	**8**	3
HSPKO	2.27	6	0	0	2	**4**
PKO	2.70	3	0	0	0	3
FFPKO	3.37	1	1	0	0	0

The Nemenyi post-hoc test further clarifies these findings. HSFFPKO was significantly better than the baseline PKO ($p = 0.010$) and FFPKO ($p = 2.0 \times 10^{-6}$), but it was not significantly different from HSPKO ($p = 0.273$). This result should be interpreted carefully. It shows that HS is the dominant contributor to the improvement over baseline PKO, while FF provides complementary stability and additional exploitation capability. Therefore, the contribution of FF is not that it always dominates as a stand-alone mechanism, but that it strengthens the full HSFFPKO framework when combined with HS.

Function-level results provide more detailed evidence. On the unimodal functions, HSFFPKO obtained the best results on F1–F3, showing strong convergence precision. On the hybrid functions, HSFFPKO achieved the best result on most cases, including F11, F12, F14–F18, and F20. This is important because hybrid functions combine different base landscapes and therefore require a balance between exploration and exploitation. The main exceptions are F13 and F19, where HSPKO obtained the lowest mean error. These exceptions show that the proposed full variant is not uniformly dominant on every function, but it remains the most reliable variant across the whole benchmark set (see Table IV to Table VI).

The composition-function results reveal an important pattern. HSFFPKO remained competitive and obtained the best result on F23, F26, and F30, but HSPKO and PKO were better on some other composition functions. This suggests that the added flocking mechanism improves robustness on some complex landscapes but may occasionally introduce additional local contraction that is less beneficial when the landscape requires broader exploration. This observation is consistent with the no-free-lunch principle and supports a cautious interpretation of the ablation results: HSFFPKO provides the best overall balance among the tested variants, although some specific functions may favor simpler or HS-only search behavior.

B. Scalability Analysis at $D = 30$, $D = 50$, and $D = 100$

To further address the scalability of the proposed mechanisms, the four PKO-based variants were evaluated on higher-

TABLE IV. ABLATION STUDY ON $D = 10$ – UNIMODAL FUNCTIONS (F1–F4).

Func.	Algorithm	Mean ± Std	Error
F1	PKO	17417.6 ± 56536.6	17317.6
	HSPKO	7308.75 ± 10802.6	7208.75
	FFPKO	5818.79 ± 4072.93	5718.79
	HSFFPKO	**1052.16** ± 1116.73	952.158
F2	PKO	307.622 ± 309.408	107.622
	HSPKO	609.48 ± 1292.54	409.48
	FFPKO	210.575 ± 28.5371	10.5753
	HSFFPKO	**200.012** ± 0.062968	0.0119784
F3	PKO	315.799 ± 40.822	15.7985
	HSPKO	304.669 ± 9.56007	4.66858
	FFPKO	300.004 ± 0.0158136	0.00395122
	HSFFPKO	**300** ± 1.14421e-06	2.12218e-06
F4	PKO	405.349 ± 1.34174	5.34882
	HSPKO	405.613 ± 0.87472	5.61319
	FFPKO	**405.049** ± 1.86189	5.0488
	HSFFPKO	405.551 ± 0.645628	5.55107

TABLE V. ABLATION STUDY ON $D = 10$ – HYBRID FUNCTIONS (F11–F20).

Func.	Algorithm	Mean ± Std	Error
F11	PKO	1104.03 ± 1.37893	4.03043
	HSPKO	1104.04 ± 1.73024	4.03600
	FFPKO	1116.72 ± 35.0349	16.7229
	HSFFPKO	**1103.54** ± 1.19288	3.54384
F12	PKO	50679.1 ± 13201.8	49479.1
	HSPKO	46992.4 ± 18392.6	45792.4
	FFPKO	39395.6 ± 22806.4	38195.6
	HSFFPKO	**29184.7** ± 15420.2	27984.7
F13	PKO	2179.94 ± 238.866	879.939
	HSPKO	**2090.08** ± 348.25	790.085
	FFPKO	2596.49 ± 2122.84	1296.49
	HSFFPKO	2489.22 ± 2292.19	1189.22
F14	PKO	1467.18 ± 29.3744	67.1786
	HSPKO	1469.19 ± 28.8256	69.1924
	FFPKO	1535.58 ± 90.5509	135.579
	HSFFPKO	**1452.59** ± 38.769	52.5937
F15	PKO	2594.11 ± 1542.44	1094.11
	HSPKO	2272.43 ± 1014.08	772.434
	FFPKO	3394.45 ± 2034.35	1894.45
	HSFFPKO	**1597.46** ± 77.6397	97.4553
F16	PKO	1614.09 ± 25.4375	14.0942
	HSPKO	1608.83 ± 11.2774	8.83311
	FFPKO	1646.31 ± 59.4109	46.3078
	HSFFPKO	**1601.44** ± 2.15571	1.4431
F17	PKO	1720.46 ± 11.1111	20.4567
	HSPKO	1716.91 ± 11.1617	16.9054
	FFPKO	1729.65 ± 13.2253	29.6493
	HSFFPKO	**1711.96** ± 9.53097	11.9635
F18	PKO	10529.9 ± 9266.26	8729.91
	HSPKO	7543.23 ± 4118.77	5743.23
	FFPKO	7712.72 ± 5124	5912.72
	HSFFPKO	**4669.6** ± 5438.7	2869.6
F19	PKO	2722.85 ± 1597.53	822.85
	HSPKO	**2596.03** ± 1174.8	696.033
	FFPKO	7548.22 ± 9779.13	5648.22
	HSFFPKO	3787.92 ± 2793.19	1887.92
F20	PKO	2011.44 ± 9.01475	11.435
	HSPKO	2010.85 ± 9.81497	10.8532
	FFPKO	2027.18 ± 26.0327	27.1775
	HSFFPKO	**2010.14** ± 10.6252	10.1425

dimensional CEC 2017 benchmark functions at $D = 30$, $D = 50$, and $D = 100$. This experiment was designed to determine whether the performance advantage observed in the $D = 10$ ablation study remains stable as the dimensionality of the search space increases. The same experimental protocol was used, including the search range $[-100, 100]^D$, the evaluation budget of $1000D$ function evaluations, and 30 independent runs per function.

Table VII shows that the Friedman test detected statistically significant differences among the four PKO variants at all tested dimensions. In each case, HSFFPKO obtained the best average rank. Its rank improved from 1.37 at $D = 30$ to 1.33 at $D = 50$ and 1.27 at $D = 100$, indicating that the combined Hovering Scouts and Foraging Flocks mechanisms remain effective as the dimensionality of the search space increases.

The win-count results in Table VIII further support the scalability of HSFFPKO. At $D = 30$, HSFFPKO obtained the best result on 24 out of 30 functions. At $D = 50$, it remained best on 23 functions, and at $D = 100$, it achieved the best result on 25 functions. In contrast, the baseline PKO obtained two wins at $D = 30$ but no wins at $D = 50$ and $D = 100$. HSPKO and FFPKO occasionally produced the best result on a small number of functions, but neither variant matched the consistency of the full HSFFPKO.

These results indicate that the two proposed mechanisms are not only useful in low-dimensional settings but also remain effective under higher-dimensional search conditions. The Hovering Scouts mechanism helps restore diversity when the search space becomes larger, while the Foraging Flocks mechanism supports coordinated exploitation of multiple promising regions. Their combination allows HSFFPKO to maintain strong rank performance and high function-win counts as dimensionality increases.

Overall, the scalability analysis strengthens the ablation findings. The $D = 10$ experiment shows that HS and FF improve the baseline search process, while the $D = 30$, $D = 50$, and $D = 100$ experiments show that the full HSFFPKO remains reliable as the dimensionality of the problem increases. Therefore, the proposed mechanisms address not only mechanism-level improvement but also high-dimensional robustness.

C. Comparison with Recent Optimizers at $D = 30$

The broader comparison evaluated HSFFPKO against nine recent metaheuristic optimizers: BWSMA, GPSOM, HMS-GWO, MLSO [25], MMRFO, OMWOA, QSO, RCCO, and SSO. The compared methods and their main design features are summarized in Table IX. This comparison was conducted at $D = 30$ to evaluate the competitiveness of HSFFPKO under a more demanding dimensional setting than the ablation study.

The Friedman test across the 30 CEC 2017 functions yielded $\chi^2 = 138.92$ with $p = 1.727 \times 10^{-25}$, confirming that the rank differences among the compared optimizers are statistically significant. As shown in Table X, HSFFPKO obtained the best overall average rank of 2.60 and the highest number of

TABLE VI. ABLATION STUDY ON $D = 10$ – COMPOSITION FUNCTIONS (F21–F30).

Func.	Algorithm	Mean ± Std	Error
F21	PKO	2270.13 ± 53.5929	170.13
	HSPKO	**2251.16** ± 54.7097	151.161
	FFPKO	2267.15 ± 58.6099	167.149
	HSFFPKO	2257.36 ± 53.4223	157.357
F22	PKO	2296.84 ± 18.291	96.8393
	HSPKO	**2294.76** ± 21.4413	94.7557
	FFPKO	2295.58 ± 20.0678	95.5791
	HSFFPKO	2300.11 ± 0.210192	100.115
F23	PKO	2611.98 ± 3.56988	311.976
	HSPKO	2610.38 ± 3.03165	310.385
	FFPKO	2619.95 ± 7.78134	319.953
	HSFFPKO	**2607.2** ± 2.30879	307.202
F24	PKO	2733.76 ± 44.3326	333.76
	HSPKO	**2721.83** ± 75.146	321.828
	FFPKO	2742.25 ± 46.4699	342.245
	HSFFPKO	2735.41 ± 2.64616	335.408
F25	PKO	**2909.41** ± 20.3816	409.406
	HSPKO	2912.64 ± 21.8685	412.643
	FFPKO	2909.98 ± 20.5569	409.977
	HSFFPKO	2926.68 ± 23.3775	426.678
F26	PKO	2916.01 ± 28.4976	316.008
	HSPKO	2910.85 ± 23.2286	310.852
	FFPKO	2948.6 ± 34.5955	348.602
	HSFFPKO	**2901.57** ± 8.60817	301.572
F27	PKO	3089.62 ± 0.385193	389.616
	HSPKO	**3089.54** ± 0.5413	389.538
	FFPKO	3090.47 ± 1.43742	390.469
	HSFFPKO	3089.56 ± 0.393655	389.555
F28	PKO	**3169.12** ± 64.1953	369.123
	HSPKO	3170.41 ± 47.6545	370.407
	FFPKO	3188.87 ± 72.3728	388.869
	HSFFPKO	3176.17 ± 96.8557	376.171
F29	PKO	**3147.1** ± 11.3148	247.103
	HSPKO	3147.85 ± 14.2279	247.854
	FFPKO	3169.52 ± 27.3461	269.521
	HSFFPKO	3148.92 ± 11.3444	248.915
F30	PKO	484720 ± 194045	481720
	HSPKO	435978 ± 202781	432978
	FFPKO	368805 ± 242669	365805
	HSFFPKO	**339458** ± 240082	336458

TABLE VII. FRIEDMAN TEST RESULTS FOR SCALABILITY EVALUATION OF PKO VARIANTS

D	Friedman χ^2	p-value	Best algorithm	Best rank
30	35.64	8.92×10^{-8}	HSFFPKO	1.37
50	34.64	1.45×10^{-7}	HSFFPKO	1.33
100	39.76	1.20×10^{-8}	HSFFPKO	1.27

function wins, with 11 wins out of 30 functions. GPSOM and BWSMA were the closest competitors, with average ranks of 2.93 and 3.00, respectively. BWSMA achieved eight function wins, while GPSOM achieved six.

The category-level results show that HSFFPKO was the only algorithm with wins across all four benchmark families. It achieved one win on unimodal functions, three wins on multimodal functions, three wins on hybrid functions, and four wins on composition functions. This distribution suggests that HSFFPKO is not specialized for only one type of landscape. Instead, its strength comes from consistent performance across different search conditions. BWSMA was particularly strong on hybrid functions, while GPSOM was highly competitive on hybrid and composition functions. These results indicate that the leading algorithms form a close top-performing group rather than a single method dominating all functions.

The function-level results in Table XI, Table XII, and Table XIII support this interpretation. HSFFPKO achieved the best result on representative functions such as F1, F21, F25, F27, and F29, and remained competitive on several other functions. BWSMA performed best on functions such as F5, F23, and F24, while GPSOM produced the best results on F22 and F30. Therefore, the comparative results show that HSFFPKO has the strongest overall rank and broadest category coverage, but the performance advantage is function-dependent.

The representative numerical results show that HSFFPKO often combines low mean error with competitive standard deviation. However, the results also show that the advantage is not uniform across all functions. Some functions favor BWSMA, GPSOM, MLSO, or QSO. This confirms that HSFFPKO should be interpreted as a strong and balanced optimizer across the suite rather than as an algorithm that dominates every individual test case.

Table XIV gives the Nemenyi post-hoc comparison between HSFFPKO and each competitor. The results show that HSFFPKO is not statistically significantly different from BWSMA, GPSOM, and MLSO at the adopted significance level. In contrast, HSFFPKO is significantly better than QSO, MMRFO, HMSGWO, SSO, RCCO, and OMWOA. This distinction is important. HSFFPKO should not be described as universally superior to all competing algorithms. A more accurate conclusion is that HSFFPKO achieved the best overall rank and function-win count, while remaining statistically comparable to the strongest competitors, especially BWSMA and GPSOM.

Thus, the supported claim is that HSFFPKO is a top-tier and well-balanced optimizer for the tested CEC 2017 setting. Its main advantage is not absolute dominance on every function, but strong and stable behavior across unimodal, multimodal, hybrid, and composition landscapes.

D. Historical Search Behavior

To further examine the search dynamics of HSFFPKO, historical search visualizations were generated for four representative CEC 2017 functions at $D = 30$: F1 as a unimodal function, F5 as a multimodal function, F13 as a hybrid function, and F23 as a composition function. For each function, the visualization includes the landscape view, the projected search distribution, the relative discrepancy curve, and the mean fitness curve. These plots provide qualitative evidence of how the population moves, contracts, and refines solutions during the optimization process.

Fig. 1 shows the behavior of HSFFPKO on the unimodal function F1. The population rapidly moves toward the optimal region, and the discrepancy curve decreases smoothly. The mean fitness curve also drops sharply in the early iterations before gradually flattening, indicating fast convergence followed

TABLE VIII. SUMMARY OF PKO-VARIANT PERFORMANCE ACROSS HIGHER DIMENSIONS ON THE CEC 2017 BENCHMARK SUITE

Algorithm	$D = 30$		$D = 50$		$D = 100$	
	Avg. rank	Wins	Avg. rank	Wins	Avg. rank	Wins
HSFFPKO	**1.37**	**24**	**1.33**	**23**	**1.27**	**25**
HSPKO	2.67	2	2.67	4	2.60	2
PKO	2.67	2	2.87	0	2.93	0
FFPKO	3.30	2	3.13	3	3.20	3

TABLE IX. COMPARED ALGORITHMS AND THE MAIN DESIGN FEATURES EMPHASIZED IN THE EXPERIMENTS

Algorithm	Core search idea	Key settings
HSFFPKO	PKO augmented with Hovering Scouts and Foraging Flocks for local diversity and group-level exploitation	scout rate 0.10, scout bias 0.05, five flocks, flocking every five iterations, exploitation weight 0.60, radius fraction 0.10, radius decay 0.95
BWSMA	best–worst population partitioning with adaptive greedy and stagnation-reset mechanisms	reset probability 0.03, best/worst grouping, covariance-based dominant-set search
GPSOM	structured multi-swarm optimizer with explicit exploration, exploitation, and equilibrium blocks	three sub-swarms (20%/20%/60%) with time-varying exploration and exploitation schedules
HMSGWO	Grey Wolf variant with hierarchical approval, omega radius, and alpha/beta schedules	canonical GWO control $a(t)$ plus multiple adaptive coefficients
MLSO	lion-inspired search with mutation and roulette-like interaction	mutation probability decays from 0.25 to 0.02 with local king-step scaling
MMRFO	manta-ray based foraging with chain, cyclone, mating, and $\mu + \lambda$ selection	stochastic chain/cyclone steps with mating-based offspring generation
OMWOA	Whale Optimization with outpost scouting and multi-population organization	WOA coefficient schedule, outpost perturbation, axis nudges, multiple subpopulations
QSO	quokka-inspired search driven by environmental and drought parameters	environment factors and exploitation mixing tied to normalized time t/T
RCCO	red-crowned crane strategy switching between foraging, roosting, and dance behaviors	strategy toggle probability with time-varying long-distance and roost steps
SSO	salp-based search strengthened by stagnation handling and local search	SSA control schedule, stagnation trigger, and local recovery attempts

TABLE X. COMPARATIVE SUMMARY ON CEC 2017 AT $D = 30$

Algorithm	Avg. rank	Wins	U	M	H	C
HSFFPKO	**2.60**	**11**	**1**	**3**	3	**4**
GPSOM	2.93	6	**1**	0	3	2
BWSMA	3.00	8	0	2	**4**	2
MLSO	4.87	2	1	0	0	1
QSO	5.37	2	1	0	0	1
MMRFO	5.67	0	0	0	0	0
HMSGWO	6.63	1	0	1	0	0
SSO	7.20	0	0	0	0	0
RCCO	8.10	0	0	0	0	0
OMWOA	8.63	0	0	0	0	0

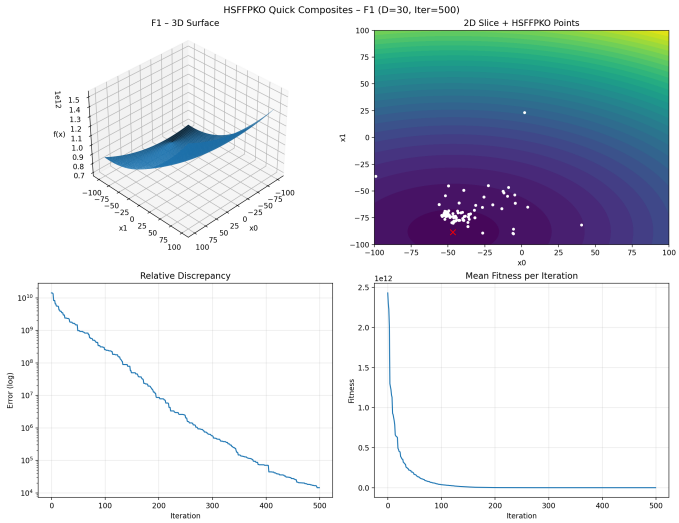


Fig. 1. Historical search on the unimodal function F1 at $D = 30$.

by local refinement. This behavior is expected for a unimodal landscape, where the main requirement is convergence speed and numerical precision.

Fig. 2 presents the behavior on the multimodal function F5. Compared with F1, convergence is slower and more stepwise because the landscape contains many local optima. The search points remain more dispersed during the earlier iterations, which suggests that the algorithm preserves exploration before committing to a promising basin. The subsequent reduction in discrepancy and mean fitness indicates that HSFFPKO can gradually escape poorer regions and refine the population around better solutions.

Fig. 3 shows the search behavior on the hybrid function F13. The population remains partially dispersed, reflecting the difficulty of optimizing a function composed of different landscape components. The mean fitness improves rapidly at first, while the discrepancy curve decreases more gradually. This suggests that HSFFPKO can identify promising regions early, but final refinement is more difficult when several landscape structures interact.

Fig. 4 illustrates the behavior on the composition function F23. The population moves toward the promising region, but some dispersion remains because the function contains shifted

TABLE XI. COMPARATIVE RESULTS ON CEC 2017 FUNCTIONS ($D = 30$) – F1–F5.

Func.	Algorithm	Mean ± Std	Error
F1	HSFFPKO	**2.002e+03** ± 1.746e+03	1.902e+03
	MLSO	1.152e+06 ± 1.512e+06	1.152e+06
	MMRFO	1.575e+06 ± 1.523e+06	1.575e+06
	OMWOA	3.839e+08 ± 4.157e+08	3.839e+08
	QSO	3.177e+05 ± 1.242e+05	3.176e+05
	RCCO	3.430e+07 ± 7.328e+07	3.430e+07
	SSO	5.002e+04 ± 5.853e+04	4.992e+04
	BWSMA	1.350e+04 ± 4.262e+04	1.340e+04
	GPSOM	6.277e+03 ± 5.774e+03	6.177e+03
	HMSGWO	1.911e+05 ± 4.127e+04	1.910e+05
F2	HSFFPKO	1.058e+13 ± 2.861e+13	1.058e+13
	MLSO	5.088e+12 ± 1.630e+13	5.088e+12
	MMRFO	1.074e+16 ± 2.339e+16	1.074e+16
	OMWOA	1.623e+21 ± 8.405e+21	1.623e+21
	QSO	**5.010e+11** ± 1.325e+12	5.010e+11
	RCCO	3.590e+21 ± 1.863e+22	3.590e+21
	SSO	3.691e+19 ± 1.798e+20	3.691e+19
	BWSMA	3.451e+13 ± 1.106e+14	3.451e+13
	GPSOM	1.435e+13 ± 7.097e+13	1.435e+13
	HMSGWO	7.198e+16 ± 1.208e+17	7.198e+16
F3	HSFFPKO	3.604e+04 ± 9.288e+03	3.574e+04
	MLSO	**5.786e+02** ± 2.966e+02	2.786e+02
	MMRFO	1.026e+04 ± 3.383e+03	9.959e+03
	OMWOA	2.823e+04 ± 8.063e+03	2.793e+04
	QSO	1.095e+04 ± 4.654e+03	1.065e+04
	RCCO	9.166e+03 ± 4.978e+03	8.866e+03
	SSO	2.065e+04 ± 7.625e+03	2.035e+04
	BWSMA	2.896e+03 ± 2.087e+03	2.596e+03
	GPSOM	1.057e+04 ± 4.663e+03	1.027e+04
	HMSGWO	3.600e+03 ± 1.692e+03	3.300e+03
F4	HSFFPKO	5.068e+02 ± 1.330e+01	1.068e+02
	MLSO	4.941e+02 ± 3.230e+01	9.408e+01
	MMRFO	5.242e+02 ± 3.759e+01	1.242e+02
	OMWOA	5.484e+02 ± 4.310e+01	1.484e+02
	QSO	5.005e+02 ± 2.071e+01	1.005e+02
	RCCO	5.186e+02 ± 2.675e+01	1.186e+02
	SSO	5.352e+02 ± 2.477e+01	1.352e+02
	BWSMA	5.060e+02 ± 2.797e+01	1.060e+02
	GPSOM	**4.866e+02** ± 2.586e+01	8.656e+01
	HMSGWO	5.201e+02 ± 2.557e+01	1.201e+02
F5	HSFFPKO	5.440e+02 ± 1.312e+01	4.402e+01
	MLSO	6.506e+02 ± 2.919e+01	1.506e+02
	MMRFO	6.602e+02 ± 3.029e+01	1.602e+02
	OMWOA	6.752e+02 ± 3.847e+01	1.752e+02
	QSO	6.504e+02 ± 3.350e+01	1.504e+02
	RCCO	7.345e+02 ± 6.103e+01	2.345e+02
	SSO	6.924e+02 ± 4.331e+01	1.924e+02
	BWSMA	**5.431e+02** ± 1.128e+01	4.308e+01
	GPSOM	5.858e+02 ± 1.781e+01	8.580e+01
	HMSGWO	6.237e+02 ± 2.059e+01	1.237e+02

TABLE XII. COMPARATIVE RESULTS ON CEC 2017 FUNCTIONS ($D = 30$) – F21–F25.

Func.	Algorithm	Mean ± Std	Error
F21	HSFFPKO	**2.343e+03** ± 9.733e+00	2.430e+02
	MLSO	2.444e+03 ± 3.062e+01	3.439e+02
	MMRFO	2.439e+03 ± 2.969e+01	3.392e+02
	OMWOA	2.468e+03 ± 3.726e+01	3.679e+02
	QSO	2.463e+03 ± 6.412e+01	3.632e+02
	RCCO	2.532e+03 ± 5.973e+01	4.321e+02
	SSO	2.485e+03 ± 4.375e+01	3.848e+02
	BWSMA	2.345e+03 ± 8.588e+00	2.452e+02
	GPSOM	2.368e+03 ± 1.755e+01	2.679e+02
	HMSGWO	2.419e+03 ± 3.571e+01	3.192e+02
F22	HSFFPKO	3.574e+03 ± 1.884e+03	1.374e+03
	MLSO	3.273e+03 ± 1.978e+03	1.073e+03
	MMRFO	2.951e+03 ± 1.705e+03	7.512e+02
	OMWOA	4.327e+03 ± 2.411e+03	2.127e+03
	QSO	4.366e+03 ± 1.918e+03	2.166e+03
	RCCO	4.593e+03 ± 2.636e+03	2.393e+03
	SSO	3.116e+03 ± 1.850e+03	9.160e+02
	BWSMA	3.769e+03 ± 1.463e+03	1.569e+03
	GPSOM	**2.306e+03** ± 4.093e+00	1.058e+02
	HMSGWO	3.290e+03 ± 2.003e+03	1.090e+03
F23	HSFFPKO	2.695e+03 ± 1.472e+01	3.947e+02
	MLSO	2.943e+03 ± 6.527e+01	6.434e+02
	MMRFO	2.905e+03 ± 7.519e+01	6.051e+02
	OMWOA	2.914e+03 ± 6.321e+01	6.135e+02
	QSO	2.846e+03 ± 5.182e+01	5.457e+02
	RCCO	2.935e+03 ± 1.072e+02	6.354e+02
	SSO	2.915e+03 ± 8.099e+01	6.145e+02
	BWSMA	**2.689e+03** ± 1.119e+01	3.892e+02
	GPSOM	2.771e+03 ± 3.716e+01	4.714e+02
	HMSGWO	2.979e+03 ± 1.266e+02	6.786e+02
F24	HSFFPKO	2.864e+03 ± 1.309e+01	4.644e+02
	MLSO	3.100e+03 ± 1.107e+02	6.998e+02
	MMRFO	3.040e+03 ± 7.865e+01	6.397e+02
	OMWOA	3.066e+03 ± 6.193e+01	6.658e+02
	QSO	2.971e+03 ± 4.651e+01	5.711e+02
	RCCO	3.060e+03 ± 6.946e+01	6.595e+02
	SSO	3.103e+03 ± 8.684e+01	7.034e+02
	BWSMA	**2.860e+03** ± 1.082e+01	4.603e+02
	GPSOM	2.911e+03 ± 2.004e+01	5.107e+02
	HMSGWO	3.123e+03 ± 9.966e+01	7.235e+02
F25	HSFFPKO	**2.898e+03** ± 1.408e+01	3.976e+02
	MLSO	2.909e+03 ± 2.355e+01	4.088e+02
	MMRFO	2.939e+03 ± 2.944e+01	4.388e+02
	OMWOA	2.972e+03 ± 4.229e+01	4.718e+02
	QSO	2.903e+03 ± 1.870e+01	4.031e+02
	RCCO	2.930e+03 ± 2.657e+01	4.298e+02
	SSO	2.930e+03 ± 2.675e+01	4.298e+02
	BWSMA	2.917e+03 ± 1.947e+01	4.167e+02
	GPSOM	2.902e+03 ± 1.899e+01	4.020e+02
	HMSGWO	2.947e+03 ± 2.154e+01	4.466e+02

and combined components. The staircase-like improvement in the discrepancy and mean fitness curves suggests that the algorithm continues to discover better basins during the search rather than converging immediately to a single region.

Overall, the historical search results support the quantitative findings. HSFFPKO converges efficiently on simpler landscapes and maintains enough diversity to continue improving on more deceptive functions. However, the visualizations also show that final localization remains more challenging on hybrid and composition functions, which motivates future work on adaptive parameter control or stronger terminal local search.

E. Engineering Design Problems

The engineering design experiments were conducted to determine whether the benchmark performance of HSFFPKO transfers to constrained real-world optimization problems. The test set includes continuous constrained problems, low-dimensional design problems, and one discrete integer problem. This allows the practical behavior of HSFFPKO to be evaluated under different types of engineering constraints.

As summarized in Table XV, HSFFPKO achieved the best standalone result on the pressure vessel problem and tied for the best result on the I-beam, speed reducer, spring, three-bar

TABLE XIII. COMPARATIVE RESULTS ON CEC 2017 FUNCTIONS ($D = 30$) – F26–F30.

Func.	Algorithm	Mean ± Std	Error
F26	HSFFPKO	4.122e+03 ± 2.045e+02	1.522e+03
	MLSO	5.818e+03 ± 1.685e+03	3.218e+03
	MMRFO	5.706e+03 ± 1.636e+03	3.106e+03
	OMWOA	5.953e+03 ± 1.121e+03	3.353e+03
	QSO	**3.705e+03** ± 1.077e+03	1.105e+03
	RCCO	5.629e+03 ± 1.869e+03	3.029e+03
	SSO	5.908e+03 ± 1.824e+03	3.308e+03
	BWSMA	3.902e+03 ± 2.575e+02	1.302e+03
	GPSOM	4.543e+03 ± 1.133e+03	1.943e+03
	HMSGWO	5.945e+03 ± 1.768e+03	3.345e+03
F27	HSFFPKO	**3.215e+03** ± 1.164e+01	5.155e+02
	MLSO	3.308e+03 ± 6.936e+01	6.084e+02
	MMRFO	3.300e+03 ± 5.981e+01	5.998e+02
	OMWOA	3.332e+03 ± 7.076e+01	6.320e+02
	QSO	3.273e+03 ± 3.527e+01	5.734e+02
	RCCO	3.295e+03 ± 6.296e+01	5.948e+02
	SSO	3.331e+03 ± 7.868e+01	6.314e+02
	BWSMA	3.233e+03 ± 1.399e+01	5.331e+02
	GPSOM	3.268e+03 ± 2.717e+01	5.685e+02
	HMSGWO	3.497e+03 ± 8.938e+01	7.969e+02
F28	HSFFPKO	3.237e+03 ± 2.193e+01	4.373e+02
	MLSO	**3.217e+03** ± 1.811e+01	4.170e+02
	MMRFO	3.272e+03 ± 3.604e+01	4.724e+02
	OMWOA	3.324e+03 ± 3.746e+01	5.240e+02
	QSO	3.227e+03 ± 2.124e+01	4.273e+02
	RCCO	3.276e+03 ± 3.275e+01	4.758e+02
	SSO	3.292e+03 ± 3.555e+01	4.915e+02
	BWSMA	3.252e+03 ± 2.996e+01	4.519e+02
	GPSOM	3.222e+03 ± 2.631e+01	4.225e+02
	HMSGWO	3.293e+03 ± 3.350e+01	4.935e+02
F29	HSFFPKO	**3.509e+03** ± 9.501e+01	6.085e+02
	MLSO	4.143e+03 ± 2.108e+02	1.243e+03
	MMRFO	4.208e+03 ± 2.453e+02	1.308e+03
	OMWOA	4.480e+03 ± 3.005e+02	1.580e+03
	QSO	4.036e+03 ± 2.376e+02	1.136e+03
	RCCO	4.496e+03 ± 4.185e+02	1.596e+03
	SSO	4.089e+03 ± 2.949e+02	1.189e+03
	BWSMA	3.561e+03 ± 1.725e+02	6.607e+02
	GPSOM	4.007e+03 ± 2.400e+02	1.107e+03
	HMSGWO	4.433e+03 ± 2.356e+02	1.533e+03
F30	HSFFPKO	1.916e+04 ± 3.481e+04	1.616e+04
	MLSO	1.897e+04 ± 3.760e+04	1.597e+04
	MMRFO	3.449e+04 ± 4.007e+04	3.149e+04
	OMWOA	2.615e+06 ± 4.927e+06	2.612e+06
	QSO	9.322e+06 ± 7.515e+06	9.319e+06
	RCCO	4.720e+05 ± 1.606e+06	4.690e+05
	SSO	4.644e+04 ± 3.933e+04	4.344e+04
	BWSMA	4.964e+04 ± 6.979e+04	4.664e+04
	GPSOM	**1.408e+04** ± 1.741e+04	1.108e+04
	HMSGWO	1.239e+07 ± 6.408e+06	1.239e+07

truss, and tubular column problems. On the cantilever beam and welded beam problems, BWSMA produced slightly better results, although HSFFPKO remained highly competitive. The main exception is the gear train problem, where HMSGWO obtained a better objective value than HSFFPKO, likely because this benchmark has a discrete integer search structure that is less naturally aligned with the continuous perturbation and flocking operators used in HSFFPKO.

The pressure vessel result is particularly important because it indicates that HSFFPKO can be effective on nonlinear constrained continuous design problems. The tied-best results on several other engineering problems further show that the proposed algorithm is not limited to synthetic benchmark functions. However, the gear train result reveals a limitation of the current design. Since the gear train problem has a discrete integer structure, the continuous perturbation and flocking mechanisms of HSFFPKO may be less naturally suited to the

TABLE XIV. NEMENYI POST-HOC COMPARISON BETWEEN HSFFPKO AND EACH COMPETITOR AT $D = 30$.

Comparator	p-value vs. HSFFPKO
BWSMA	1.000
GPSOM	1.000
MLSO	1.056×10^{-1}
QSO	1.467×10^{-2}
MMRFO	3.485×10^{-3}
HMSGWO	1.097×10^{-5}
SSO	1.791×10^{-7}
RCCO	8.924×10^{-11}
OMWOA	5.322×10^{-13}

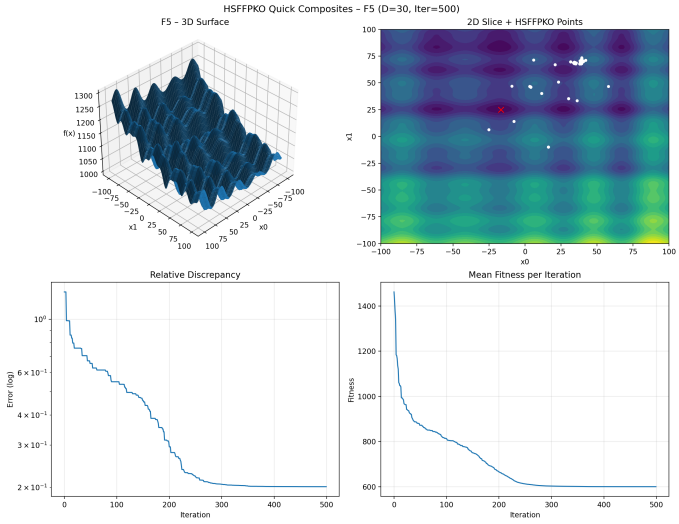


Fig. 2. Historical search on the multimodal function F5 at $D = 30$.

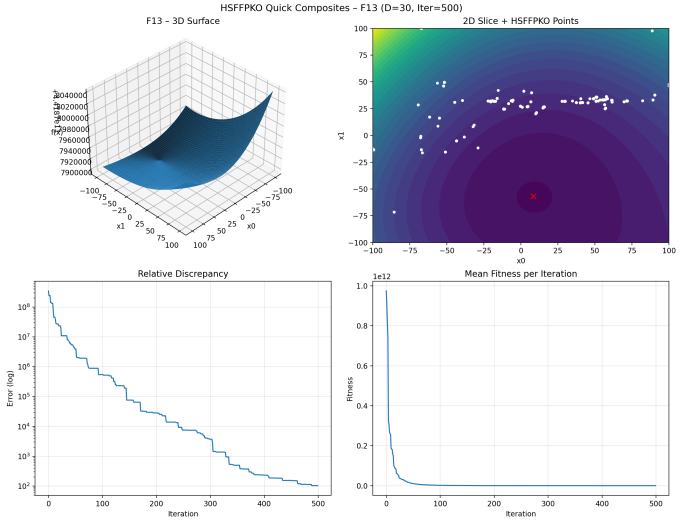


Fig. 3. Historical search on the hybrid function F13 at $D = 30$.

TABLE XV. CONDENSED ENGINEERING BENCHMARK RESULTS

Problem	Vars	Winning algorithm(s) and best mean	HSFFPKO outcome	Interpretation
Cantilever beam	5	BWSMA, 1.339956316	1.339957370; competitive	near-tie on a smooth constrained continuous problem
Gear train	4 int.	HMSGWO, 2.70×10^{-12}	5.70×10^{-11}; weaker	discrete ratio matching favors a different search dynamic
I-beam	4	GPSOM and HSFFPKO, 98.59387445	tied-best	stable convergence with negligible variance
Pressure vessel	4	HSFFPKO, 6049.858062	best overall	strongest exploitation under nonlinear constraints
Speed reducer	7	BWSMA, GPSOM, HSFFPKO, MLSO, 60503335.93	tied-best	several methods can reach the optimum on this case
Spring	3	GPSOM, HSFFPKO, RCCO, 0.001350022	tied-best	machine-precision convergence by multiple methods
Three-bar truss	2	all methods, 0.382842712	tied-best	problem is easy enough that ranking is not discriminative
Tubular column	2	BWSMA, HSFFPKO, MLSO, OMWOA, RCCO, SSO, 666.6666654	tied-best	another benchmark with many exact ties
Welded beam	4	BWSMA, 2.845786325	2.848179512; competitive	BWSMA holds a small edge on this nonlinear design

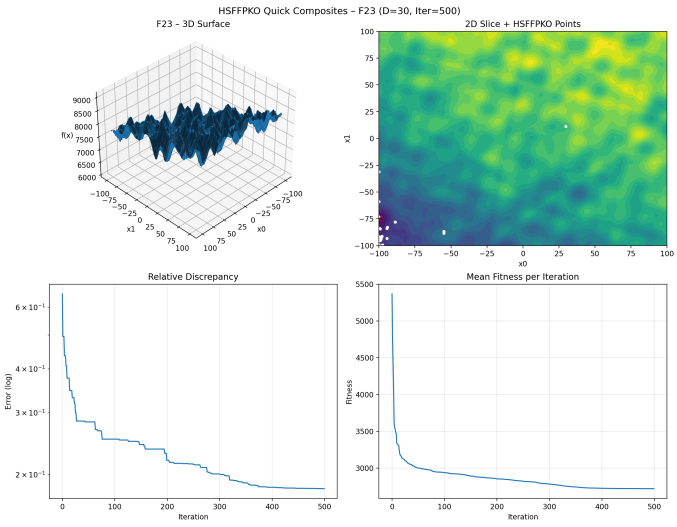


Fig. 4. Historical search on the composition function F23 at $D = 30$.

search space. In this case, the coordinated continuous basin search introduced by FF does not provide the same advantage that it provides on continuous problems.

The overall Friedman ranking on the engineering set places HSFFPKO first with an average rank of 2.67, followed closely by BWSMA with 2.72 and GPSOM with 3.72. This confirms that HSFFPKO is among the strongest methods for the engineering test set. Nevertheless, the results also support a cautious interpretation: HSFFPKO is highly competitive for continuous nonlinear constrained optimization, but specialized or more discrete-aware operators may be needed for integer or mixed-integer design problems.

F. Discussion

The results can be interpreted by viewing HSFFPKO as a layered search process. The baseline PKO already provides exploration through perching and hovering, exploitation through diving, and additional diversity through commensal association. However, the baseline method does not explicitly separate individual-level diversity recovery from subgroup-level basin exploitation. HS and FF were designed to address these two roles.

The ablation results show that HS is the dominant contributor to the performance gain. This is supported by the strong rank of HSPKO and the non-significant difference between HSPKO and HSFFPKO in the post-hoc test. The reason is that HS directly reactivates selected agents through scale-aware Gaussian probing, which helps reduce premature convergence and improves local diversity. FF, on the other hand, is more useful as a complementary mechanism because it organizes the population into temporary flocks and promotes coordinated exploitation around leader–centroid targets. The full HSFFPKO benefits from both mechanisms, but the improvement is not uniform across all function types.

The comparison with recent optimizers shows that HSFFPKO is a strong general-purpose optimizer. It achieved the best overall Friedman rank and the largest number of function wins at $D = 30$. However, the Nemenyi test shows that its differences from BWSMA and GPSOM are not statistically significant. Thus, HSFFPKO belongs to the top-performing group and offers the best overall balance in the present experiments.

G. Practical Implications

From a practical perspective, HSFFPKO has three main advantages. First, it is simple to integrate into the baseline PKO because HS and FF are inserted as additional operators without changing the main PKO structure. Second, it maintains the same asymptotic complexity class as PKO, which makes the additional search capability computationally manageable. Third, it performs well across both benchmark functions and constrained engineering problems, suggesting that it can be useful for practical optimization tasks where the search landscape is unknown.

The engineering results suggest that HSFFPKO is particularly suitable for continuous nonlinear constrained optimization problems. Its performance on the pressure vessel

problem and its tied-best results on several classical design problems indicate that the proposed mechanisms can improve practical search behavior. The algorithm is therefore promising for engineering applications where objective functions are expensive, nonlinear, or difficult to differentiate.

H. Limitations

Despite its strong overall performance, HSFFPKO has several limitations. First, although the scalability analysis shows that HSFFPKO remains effective at $D = 30$, $D = 50$, and $D = 100$, the full comparison with all recent optimizers was emphasized at $D = 30$. Future work should extend the broad multi-algorithm comparison to higher dimensions to determine whether the same relative ranking among HSFFPKO, BWSMA, GPSOM, and other competitors remains stable under larger-scale settings.

Second, the proposed mechanisms are mainly designed for continuous search spaces. The weaker result on the discrete gear train problem suggests that HSFFPKO may require discrete or mixed-integer adaptations for problems with integer variables. Third, the current parameter settings were fixed across all benchmark functions. Although this supports fairness and simplicity, adaptive control of scout rate, scout bias, flocking interval, exploitation weight, and radius decay may further improve robustness.

Finally, the historical search visualizations show that final refinement remains difficult on some hybrid and composition functions. This suggests that HSFFPKO could benefit from a stronger terminal local-search procedure, adaptive radius control, or hybridization with problem-specific refinement operators.

VII. Conclusion

This study presented HSFFPKO, an enhanced variant of PKO that integrates Hovering Scouts and Foraging Flocks to improve diversity preservation, coordinated basin search, and late-stage refinement. HS performs scale-aware local probing with a weak directional bias toward the current best solution, while FF creates temporary subgroups that search around leader–centroid targets under a shrinking radius. The proposed method preserves the simple structure of PKO while adding two complementary search mechanisms that operate at different scales.

The experimental results show that HSFFPKO provides strong and consistent performance across benchmark and engineering problems. In the ablation study at $D = 10$, HSFFPKO achieved the best average rank of 1.67 and obtained 20 wins out of 30 CEC 2017 functions. In the scalability analysis, HSFFPKO remained the best PKO-based variant at $D = 30$, $D = 50$, and $D = 100$, with average ranks of 1.37, 1.33, and 1.27, respectively. In the broad comparison with recent optimizers at $D = 30$, HSFFPKO obtained the best overall average rank of 2.60 and the largest number of function wins. However, the Nemenyi post-hoc test showed that HSFFPKO was statistically comparable with BWSMA and GPSOM, indicating that it should be interpreted as a top-tier optimizer rather than a universally dominant method.

The engineering results further confirmed the practical relevance of HSFFPKO. The proposed algorithm achieved the best result on the pressure vessel problem and tied for the best result on several other constrained design problems. Nevertheless, its weaker result on the discrete gear-train problem suggests that future work should develop discrete or mixed-integer variants of HSFFPKO. Other future directions include adaptive parameter control, stronger local exploitation on deceptive hybrid and composition functions, surrogate-assisted variants for expensive optimization, and extension to multi-objective optimization.

References

[1] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 – International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995, doi: 10.1109/ICNN.1995.488968.

[2] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[3] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016, doi: 10.1016/j.advengsoft.2016.01.008.

[4] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Advances in Engineering Software*, vol. 114, pp. 163–191, 2017, doi: 10.1016/j.advengsoft.2017.07.002.

[5] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, Studies in Computational Intelligence, vol. 284, J. R. González, D. A. Pelta, C. Cruz, G. Terrazas, and N. Krasnogor, Eds. Berlin, Heidelberg: Springer, 2010, pp. 65–74, doi: 10.1007/978-3-642-12538-6_6.

[6] J. Kang and Z. Ma, "Red-Crowned Crane Optimization: A novel biomimetic metaheuristic algorithm for engineering applications," *Biomimetics*, vol. 10, no. 9, p. 565, 2025, doi: 10.3390/biomimetics10090565.

[7] W. J. Al-Kubaisy and B. Al-Khateeb, "Quokka swarm optimization: A new nature-inspired metaheuristic optimization algorithm," *Journal of Intelligent Systems*, vol. 33, no. 1, 2024, doi: 10.1515/jisys-2024-0051.

[8] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997, doi: 10.1109/4235.585893.

[9] D. Farinati and L. Vanneschi, "A survey on dynamic populations in bio-inspired algorithms," *Genetic Programming and Evolvable Machines*, vol. 25, no. 2, 2024, doi: 10.1007/s10710-024-09492-4.

[10] Z. Jakšić, S. Devi, O. Jakšić, and K. Guha, "A comprehensive review of bio-inspired optimization algorithms including applications in microelectronics and nanophotonics," *Biomimetics*, vol. 8, no. 3, p. 278, 2023, doi: 10.3390/biomimetics8030278.

[11] M. Sarhani, S. Voß, and R. Jovanovic, "Initialization of metaheuristics: Comprehensive review, critical analysis, and research directions," *International Transactions in Operational Research*, vol. 30, no. 6, pp. 3361–3397, 2022, doi: 10.1111/itor.13237.

[12] W. H. Bangyal, J. Ahmed, and H. T. Rauf, "A modified bat algorithm with torus walk for solving global optimisation problems," *International Journal of Bio-Inspired Computation*, vol. 15, no. 1, pp. 1–13, 2020, doi: 10.1504/IJBIC.2020.105861.

[13] W. H. Bangyal, A. Hameed, J. Ahmad, K. Nisar, M. R. Haque, A. A. A. Ibrahim, J. J. P. C. Rodrigues, M. A. Khan, D. B. Rawat, and R. Etengu, "New modified controlled bat algorithm for numerical optimization problem," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2241–2259, 2022, doi: 10.32604/cmc.2022.017789.

[14] W. H. Bangyal, J. Ahmad, and H. T. Rauf, "Optimization of neural network using improved bat algorithm for data classification," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 4, pp. 670–681, 2019, doi: 10.1166/jmihi.2019.2654.

[15] A. Bouaouda, F. A. Hashim, Y. Sayouti, and A. G. Hussien, "Pied kingfisher optimizer: A new bio-inspired algorithm for solving numerical optimization and industrial engineering problems," *Neural Computing and Applications*, 2024, doi: 10.1007/s00521-024-09879-5.

[16] W. Zhao, Z. Zhang, and L. Wang, "Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103300, 2020, doi: 10.1016/j.engappai.2019.103300.

[17] A. N. K. Nasir, M. A. A. Roslan, M. F. M. Jusof, M. R. Ahmad, and A. A. A. Razak, "Mating-based manta ray foraging optimization for fuzzy-Hammerstein model of an electric water heater," *Journal of Advanced Research in Applied Mechanics*, vol. 129, no. 1, pp. 32–43, 2024, doi: 10.37934/aram.129.1.3243.

[18] K. Tang and L. Zhang, "An enhanced whale optimization algorithm with outpost and multi-population mechanisms for high-dimensional optimization and medical diagnosis," *PLOS ONE*, vol. 20, no. 6, p. e0325272, 2025, doi: 10.1371/journal.pone.0325272.

[19] T. Li, H. Meng, D. Wang, B. Fu, Y. Shao, and Z. Liu, "An enhanced slime mould algorithm based on best–worst management for numerical optimization problems," *Biomimetics*, vol. 10, no. 8, p. 504, 2025, doi: 10.3390/biomimetics10080504.

[20] J. Yan, G. Hu, H. Jia, A. G. Hussien, and L. Abualigah, "GPSOM: Group-based particle swarm optimization with multiple strategies for engineering applications," *Journal of Big Data*, vol. 12, article 114, 2025, doi: 10.1186/s40537-025-01140-7.

[21] I. Dagal, A. W. Ibrahim, A. Harrison, *et al.*, "Hierarchical multi step Gray Wolf optimization algorithm for energy systems optimization," *Scientific Reports*, vol. 15, article 8973, 2025, doi: 10.1038/s41598-025-92983-w.

[22] N. H. Awad, M. Z. Ali, P. N. Suganthan, J. J. Liang, and B. Y. Qu, "Problem definitions and evaluation criteria for the CEC 2017 special session and competition on single objective real-parameter numerical optimization," Technical Report, Nanyang Technological University, Singapore, 2016.

[23] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011, doi: 10.1016/j.swevo.2011.02.002.

[24] E. H. Houssein, M. H. A. Gafar, N. Fawzy, and A. Y. Sayed, "Recent metaheuristic algorithms for solving some civil engineering optimization problems," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-90000-8.

[25] F. Ji and M. Jiang, "Lion swarm optimization by reinforcement pattern search," in *Advances in Swarm Intelligence*, Lecture Notes in Computer Science, vol. 12689. Cham: Springer, 2021, pp. 119–129, doi: 10.1007/978-3-030-78743-1_11.

A Lightweight Explainable Hybrid Deep Learning Approach for Early Skin Disease Detection

Mohammad Barr
Department of Electrical Engineering, College of Engineering
Northern Border University, Arar 91431, Saudi Arabia

***Abstract*—Early detection of skin pathology is critical for patient survival and treatment effectiveness, particularly in the case of aggressive malignancies such as melanoma. In this study, we propose a lightweight and explainable hybrid deep learning approach for early identification of skin diseases. The proposed approach combines a transformer-based module to capture the global contextual dependencies with a CNN for an efficient local feature extraction. In addition, Grad-CAM approach is used to increase the interpretability of the model and offer visual explanations for the forecasts. We evaluate the suggested technique on a publically accessible dermoscopic benchmark and achieve 93% accuracy, 92% precision, 91% recall and 92% F1-score, outperforming numerous mainstream designs. The experimental results imply that the proposed approach achieves a good trade-off between accuracy, computational economy, and interpretability for real-world and resource-limited medical applications.**

***Keywords*—Medical image analysis; skin disease detection; explainable AI; hybrid model; Vision Transformer**

I. Introduction

Skin diseases are one of the most common health problems in the world, affecting millions of people every year. This can be anything from benign acne and eczema to serious melanoma. Early detection is very crucial, especially for malignant skin cancers because it increases the survival rates and decreases the treatment costs. However, the accurate diagnosis of these conditions is still a challenge due to the similarities between different skin lesions and variation in the imaging conditions such as lighting, resolution, and skin tone [1].

Until recently, the clinical diagnosis of skin diseases has been based on the clinical examination and the dermoscopic examination of an experienced expert. This technique, while successful, is fundamentally subjective and relies on clinical experience, which may result in inter-observer variability and delayed diagnosis. Moreover, the number of dermatologists is not enough in many places, particularly in developing nations, generating a tremendous need for automated, scalable, and reliable diagnostic solutions [1], [2].

Recently, deep learning has been successfully used in medical image analysis, in particular, in the detection of skin diseases. CNNs have demonstrated themselves to be quite effective in extracting hierarchical features from the pictures to appropriately identify the skin lesions. A lot of research has demonstrated a significant increase in the accuracy of diagnosis by using deep learning algorithms ([2], [3]). These models automatically learn discriminative features like texture, color distribution, and the borders of lesions without the need for handcrafted feature engineering.

CNN-based techniques have been successful, but suffer from several limitations. In particular, CNNs are meant to capture local spatial characteristics by using convolution processes with a narrow receptive field. Deeper designs may partly overcome this constraint but generally fail to simulate long-range relationships inside images. This is especially important in medical imaging because small global patterns may be important for distinguishing different disease categories [3].

Recently, transformer-based architectures have been used for computer vision applications to solve these problems. After their success in natural language processing, Transformers use self-attention processes to model connections between all elements in the incoming data. Vision Transformer (ViT) models split pictures into patches and treat them as sequences to collect global contextual information. Recent research has shown that transformer-based models are useful for classifying skin diseases, and they achieve better performance than typical CNNs [4], [5].

However, transformer models also have certain problems. They are computationally complicated and need large-scale datasets for efficient training, which makes them less suited for real-world and resource-constrained applications. Also, CNN and transformer models typically act as black-box systems and lack interpretability, which is a critical problem in medical applications where transparency and trust are important [4].

Recently, hybrid deep learning architectures that blend CNNs and transformers have been proposed to solve the above shortcomings. These methods leverage CNNs for effective local feature extraction and transformers to model global dependencies. This combination enables the model to benefit from the benefits of both paradigms and achieve better performance and resilience. Hybrid CNN-Transformer models have shown their efficacy in skin disease diagnosis in recent research, performing better than stand-alone designs in terms of accuracy [6], [7].

In addition to speed, interpretability has become a key criterion in medical AI systems. Clinicians need to grasp the rationale behind the model's predictions to trust and use these technologies in their practice. Explainable Artificial Intelligence (XAI) solutions try to tackle this problem by shedding light on the decision-making phase of deep learning approaches. Among such approaches, Grad-CAM has drawn substantial attention for its capacity to provide visual explanations by highlighting critical locations in the input picture. Recent works successfully utilized XAI approaches to skin disease categorization, displaying enhanced transparency and clinical relevance [3], [8].

However, while these advances have been made, most of the current systems focus either on improving the accuracy or on increasing the interpretability, but few of them try to improve both while maintaining high computational efficiency. Nor has there been as much attention to early-stage identification, when signs of illness are less severe and harder to detect. This gap highlights the need for a unifying framework that encompasses accuracy, efficiency, and explainability. To address such problems, this study proposes a lightweight explainable hybrid intelligent system for early skin disease diagnosis. The proposed method consists of a CNN-based feature extractor and a transformer-based module, which can successfully collect local and global features. The architecture also has Grad-CAM for providing visual explanations of model predictions, thus enhancing interpretability and confidence.

The principal contributions of this study are listed below:

- We propose a lightweight hybrid CNN-Transformer architecture for early skin disease detection, combining MobileNetV2 for local feature extraction with a ViT for global context modeling.
- We integrate Grad-CAM, an explainable AI technique, into the hybrid framework to provide visual heatmaps that highlight lesion regions influencing model predictions.
- We optimize the proposed hybrid model for computational efficiency.
- We conduct a comprehensive evaluation on benchmark datasets with a comparative analysis.

The rest of this work is structured as follows: In Section II, we review existing approaches and recent studies related to skin disease detection. The proposed hybrid CNN-Transformer technique is discussed in Section III. The experimental setups are described in Section IV. Quantitative and qualitative data are presented and analyzed in Section V. In Section VI, we conclude the study and outline several directions for future research.

II. Related Work

The development of AI has significantly improved the ability of automated diagnostic systems for skin diseases. In particular, deep learning algorithms have achieved great success in medical image analysis owing to their capacity to automatically generate discriminative features from raw data. This section introduces the most important literature, separated into CNN-based methods, transformer-based models, hybrid architectures, and explainable AI strategies.

A. CNN-Based Methods for Skin Disease Detection

CNNs have become one of the cornerstones for automated skin lesion analysis, since they are capable of learning hierarchical spatial representations directly from raw images. These models learn fine-grained patterns well, such as textural differences, color distribution, and lesion edges [1]. More and more, current research has concentrated on the improvement of CNN performance using transfer learning and complex optimization strategies. For example, Zhang et al. [1] provided a detailed assessment of machine and deep learning techniques for dermatological diagnosis and spoke about the efficiency of CNN-based systems in achieving high discrimination accuracy. Likewise, Liu et al. [9] performed a thorough evaluation of deep learning applications in dermatology, where convolutional architectures are gradually taking over clinical procedures. Khan et al. [10] showed that transfer learning-based CNNs outperform traditional machine learning classifiers for skin lesion identification.

B. Transformer Architectures

Recently, the computer vision community has started to incorporate transformer-based models initially created for natural language processing to address the receptive field constraint of CNNs. The self-attention mechanism is at the core of these systems, which allows direct modeling of long-range dependencies and global interactions among picture patches [4]. The significance of transformers in medical imaging activities has been proved by many research. Mohan et al. [11] proposed a transformer-based deep learning pipeline for skin disease classification, which resulted in superior performance due to enhanced global context awareness. Similarly, Dagnaw et al. [12] deployed a Vision Transformer for melanoma detection and included explainability tools to enhance the interpretability of the model.

C. Hybrid CNN-Transformer Models

Recent works have been done on hybrid architectures that mix the two approaches to take use of the complementary qualities of CNNs and transformers. CNNs are utilized in such models for effective local feature extraction and transformers for capturing global contextual linkages. Ali et al. [13] introduced a hybrid CNN-Transformer architecture for skin disease identification, which produced better results than using just CNN or Transformer models. Their method underlines the need for combining local and global feature representations. Also, Mousa et al. [14] developed an attention-based hybrid CNN-Transformer model for medical picture classification, which showed better robustness and generalization. These investigations demonstrate that hybrid designs may achieve a good trade-off between accuracy and computing economy. However, even with their enhanced performance, many hybrid models still suffer from computational expense and lack of interpretability, which are crucial in medical applications.

D. Explainable AI for Skin Disease Detection

In medical AI applications, the raw prediction performance is not adequate. In order to trust and use a model in practice, clinicians need to grasp the rationale behind a given choice made by the model. This need has brought XAI to the forefront of the development of diagnostic systems. The goal of XAI approaches is to provide human-interpretable explanations of the internal decision processes of deep learning models, thereby opening the "black box" of these models. Of these approaches, Gradient-weighted Class Activation Mapping (Grad-CAM) has garnered particular momentum in dermatological imaging for its capacity to provide visual heatmaps detailing which parts of a skin lesion picture most impacted the model's output. Shah et al. [15] proposed a XAI-based framework for skin cancer identification and showed the Grad-CAMs' ability to locate diagnostic-relevant regions such as lesion boundaries,

pigmented networks, and vascular structures. Similarly, Badhon et al. [16] use Grad-CAM on multi-class skin disease classification and demonstrate that visual explanations using Grad-CAM considerably increase the transparency and clinical value of the models without any degradation in the classification accuracy.

III. Proposed Approach: Hybrid CNN–Transformer with Explainability

To detect skin diseases early, we propose a deep hybrid framework that combines CNNs with Vision Transformers in an explainable manner. This architecture efficiently captures both local spatial cues and global contextual information, yet remains computationally economical and interpretable thanks to integrated XAI. The input preprocessing, CNN-based feature extraction, transformer-based global modeling, classification head, and explainability module are the five primary components of the suggested system, as shown in Fig. 1. A lightweight CNN extracts discriminative spatial features, which are then processed by a transformer encoder to model long-range dependencies. Finally, predictions are generated and interpreted using visualization techniques.

A. Input Preprocessing and Data Augmentation

Each dermoscopic image I from the Skin Cancer (MNIST HAM10000) dataset is preprocessed to ensure input consistency and improve model generalization. All images are resized to a fixed resolution of $224\times224\times3$ pixels. Subsequently, pixel values are normalized channel-wise using the dataset mean μ and standard deviation σ:

$$I_{\text{norm}} = \frac{I - \mu}{\sigma} \tag{1}$$

The HAM10000 dataset comprises 10,000 dermoscopic images labeled into seven skin lesion categories [17]. As illustrated in Fig. 2, the class distribution is highly imbalanced: melanocytic nevi dominate with approximately 66.9% (6,705 images), while minority classes such as dermatofibroma (1.1%) and vascular lesions (1.4%) contain far fewer samples. The remaining classes include melanoma (11.1%), benign keratosis-like lesions (11%), basal cell carcinoma (5.1%), and actinic keratoses (3.3%).

Such imbalance can bias the classifier toward majority classes if left unaddressed. Therefore, we apply data augmentation to enhance model robustness and mitigate overfitting. The augmentation pipeline includes random rotations, horizontal and vertical flips, scaling, and cropping. These transformations increase the model's invariance to variations in orientation, size, and lighting conditions commonly encountered in dermoscopic practice. Moreover, augmentation helps to artificially expand the representation of minority classes, partially alleviating the class imbalance problem and leading to a more balanced and unbiased model.

The vast visual diversity and morphological complexity of skin lesions are illustrated in Fig. 3, which includes representative samples from each of the seven classes.

B. CNN-Based Feature Extraction

The normalized input image I_norm is processed using the MobileNetV2 backbone [18] [19] to produce a highly discriminative yet compact feature representation. This lightweight architecture is specifically engineered for efficiency, employing depthwise separable convolutions to significantly reduce computational complexity while preserving a robust feature extraction capability. The network begins with an initial convolutional layer, followed by a sequence of inverted residual blocks with linear bottlenecks, which form the core of the architecture. These blocks are designed to enhance feature reuse and improve gradient flow during training. Each inverted residual block can be mathematically expressed as:

$$\mathbf{y} = \mathbf{x} + \mathcal{F}(\mathbf{x}) \tag{2}$$

where, x denotes the input feature map and $F(\cdot)$ represents a series of transformations consisting of: Pointwise convolution, Depthwise convolution, and Pointwise linear projection.

This structure enables efficient feature learning by separating spatial and channel-wise computations, thereby preserving essential information while minimizing memory usage and computational cost. After passing through multiple stacked inverted residual blocks, the network produces a high-level feature map: $F \in \mathbf{R}^{7\times7\times1280}$.

To further refine the extracted features, a 1×1 convolution is applied, followed by global average pooling, which aggregates spatial information into a more compact representation. This refined feature representation serves as the input to the subsequent transformer module, enabling effective integration of local spatial features with global contextual modeling.

C. Transformer-Based Global Modeling

The proposed framework integrates a ViT-based encoder [20] to efficiently capture global contextual relationships and long-range dependencies within the extracted feature map. The feature map generated by the CNN backbone is first reshaped into a sequence of non-overlapping patches, denoted as:

$$\{x_1, x_2, \ldots, x_N\}$$

where, each patch represents a local region of the input image. Each patch is subsequently flattened and projected into a latent embedding space through a learnable linear transformation, expressed as $z_i = x_i W_e$, where W_e denotes the embedding matrix.

To preserve the spatial structure of the original image, positional embeddings are added to each token, resulting in $z_i' = z_i + p_i$, where p_i encodes the positional information. Furthermore, a special classification token is appended to the input sequence, serving as a global representation that aggregates information from all patches throughout the transformer layers.

The augmented sequence of embeddings is subsequently processed by a series of transformer encoder layers. Each encoder layer comprises a multi-head self-attention mechanism succeeded by a feed-forward neural network, incorporating layer normalization and residual connections to improve training stability and convergence. The self-attention mechanism

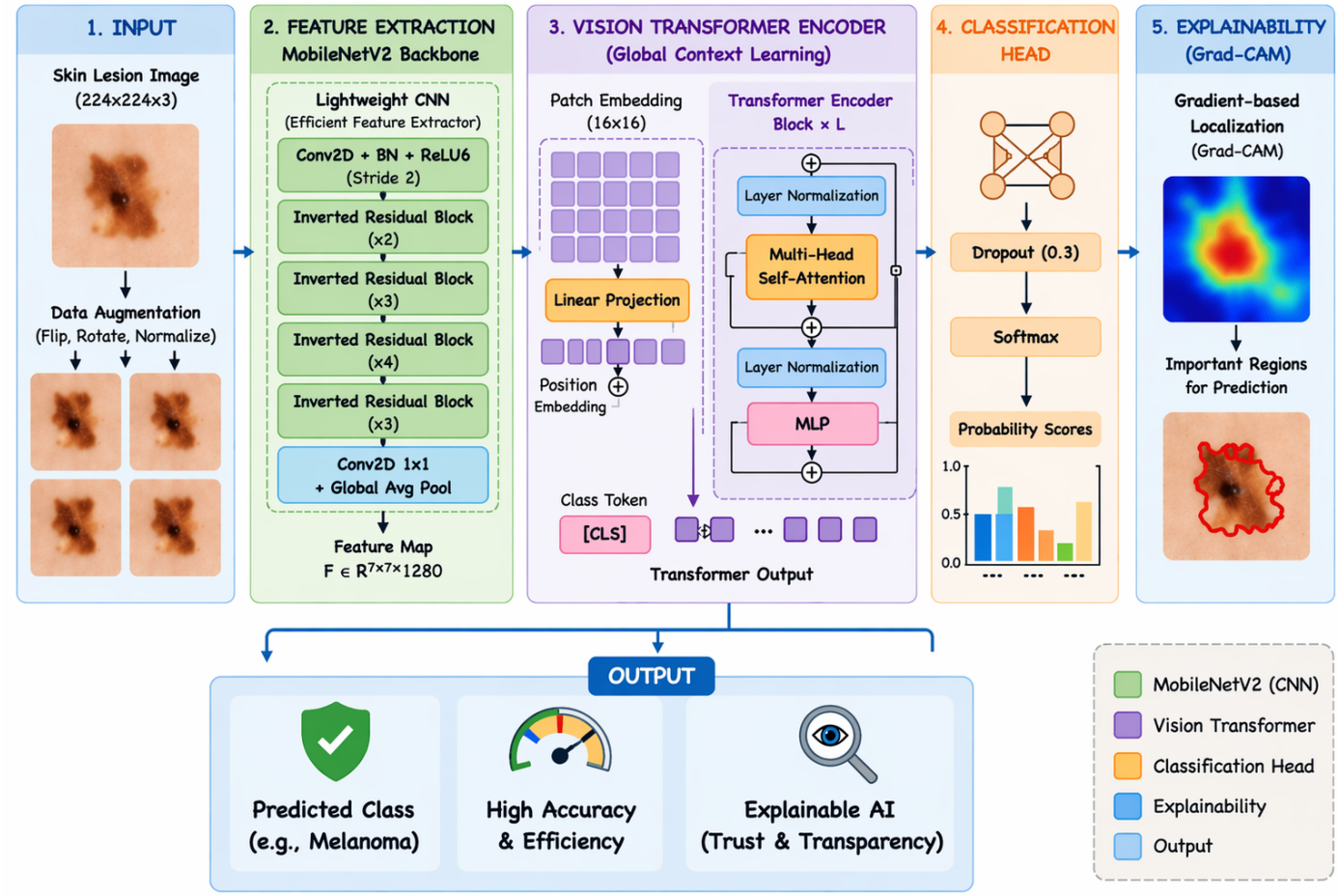


Fig. 1. Overview of the proposed framework.

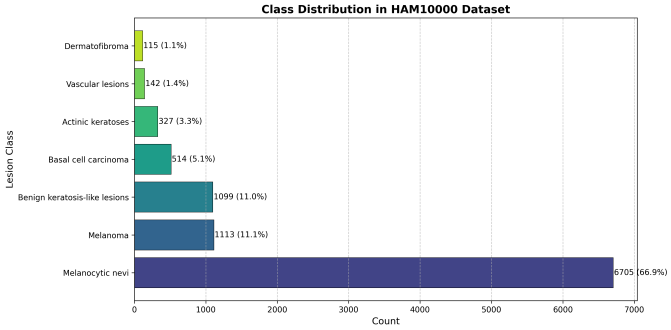


Fig. 2. Skin lesion class distribution.

allows the model to evaluate relationships among all patches, enabling it to concentrate on the most distinguishing areas of the lesion image irrespective of their spatial positioning.

Finally, the output corresponding to the classification token is extracted and used as a compact global representation of the input image. This representation encodes high-level semantic and contextual information and is forwarded to the classification head to perform the final skin disease prediction.

D. Explainability Using Grad-CAM

To make our model's decisions more transparent and interpretable, we integrate Grad-CAM. This technique produces visual heatmaps that reveal which portions of the input image most strongly influenced the model's prediction [21]. Grad-CAM works by backpropagating the gradient of the target class score to the final convolutional layer's feature maps. This process quantifies how much each feature map contributes to the classification of a particular class.

Formally, the importance weight associated with the $k-th$ feature map for class c is computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

where, A_{ij}^k represents the activation value at spatial coordinates (i,j) within the $k-th$ feature map, y^c denotes the predicted score for class c, and Z normalizes by the total number of spatial locations in the feature map.

Once these weights are obtained, we generate the Grad-CAM heatmap by taking a weighted linear combination of the feature maps, followed by a ReLU activation function to suppress negative contributions:

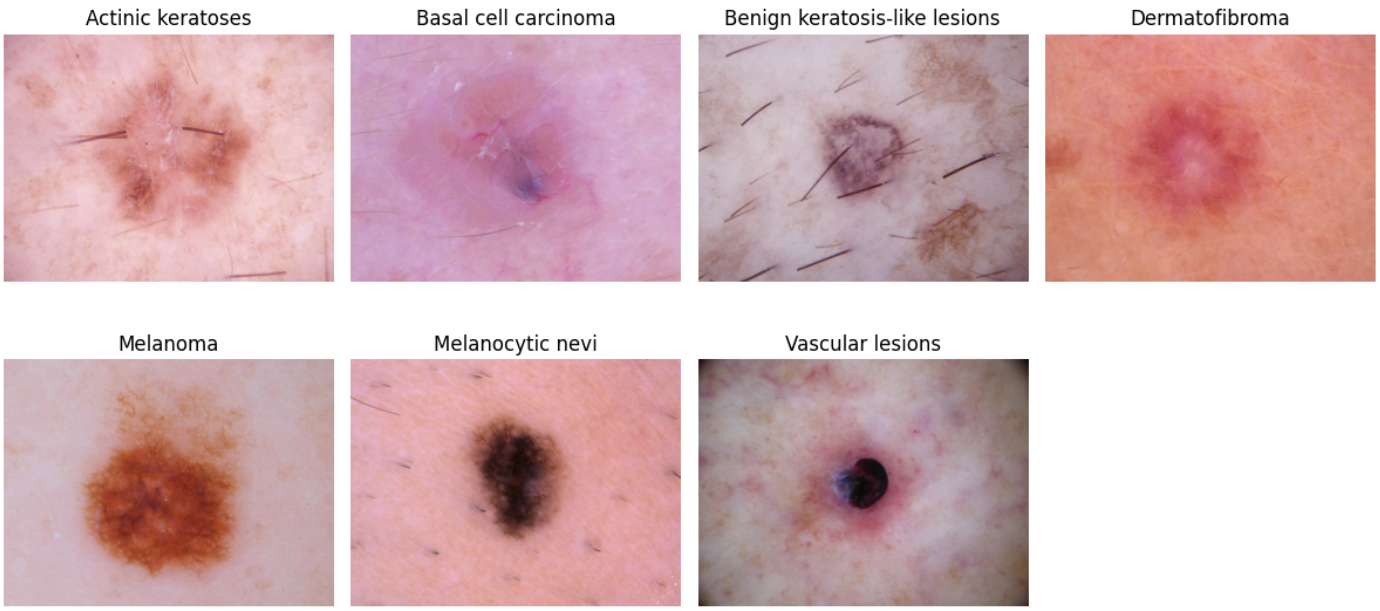


Fig. 3. Sample images from each class.

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{4}$$

In this equation, A^k refers to the $k-th$ feature map and α^c_k is its associated weight. The resulting heatmap highlights the image regions most relevant to the model's decision. This visual explanation enhances transparency and helps clinicians understand and trust the model's diagnostic reasoning.

IV. Experimental Setup

A. Dataset Description

The suggested model is evaluated on public benchmark datasets for skin disease classification [17]. Specifically, the model is validated on dermoscopic image datasets with multiple categories of skin lesions such as melanoma, vascular lesions, actinic keratoses, and benign keratosis-like lesions. The dataset consists of dermoscopic images with high resolution acquired in different settings with different lighting conditions, patient skin color, and lesion morphology. This variation helps to encourage model robustness and generalizability to real-world clinical settings. The images are labeled manually with the corresponding disease label by dermatology experts. For experimental purposes, the data is partitioned into three separate parts: 80% for training (for training the model parameters), 10% for validation (for hyperparameter tuning), and the remaining 10% as a test set for the final performance evaluation. Our architecture takes an input size of 224×224 pixels. Therefore, all images are resized to a fixed size of 224×224 pixels. Details of the dataset are described in Section III-A.

B. Evaluation Metrics

Table I provides an overview of the multiple metrics we use to thoroughly assess the proposed model's performance.

TABLE I. Experimental Metrics

Metric	Name
$Acc = \frac{TP+TN}{TP+TN+FP+FN}$	Accuracy
$P = \frac{TP}{TP+FP}$	Precision
$R = \frac{TP}{TP+FN}$	Recall
$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	F1 Score

C. Implementation Details

Our implementation relies on the TensorFlow and Keras deep learning libraries. All experiments run on GPU-accelerated hardware using Kaggle notebooks, which speeds up training and supports reproducible research. The network accepts input images of dimensions $224 \times 224 \times 3$ and processes them in batches of 32 samples. The Adam optimizer is utilized for optimization, with an initial learning rate of 10^{-4}. The categorical cross-entropy loss function is employed for multi-class classification. Training proceeds for 30 epochs, though early stopping can be applied depending on convergence patterns. The MobileNetV2 backbone is seeded with ImageNet pretrained weights, allowing the model to benefit from transfer learning for more efficient feature extraction. Meanwhile, the transformer encoder and the final classification head are trained together in an end-to-end fashion to maximize overall predictive performance.

D. Training Strategy

To ensure stable and efficient training, several strategies are adopted within the proposed framework. First, transfer

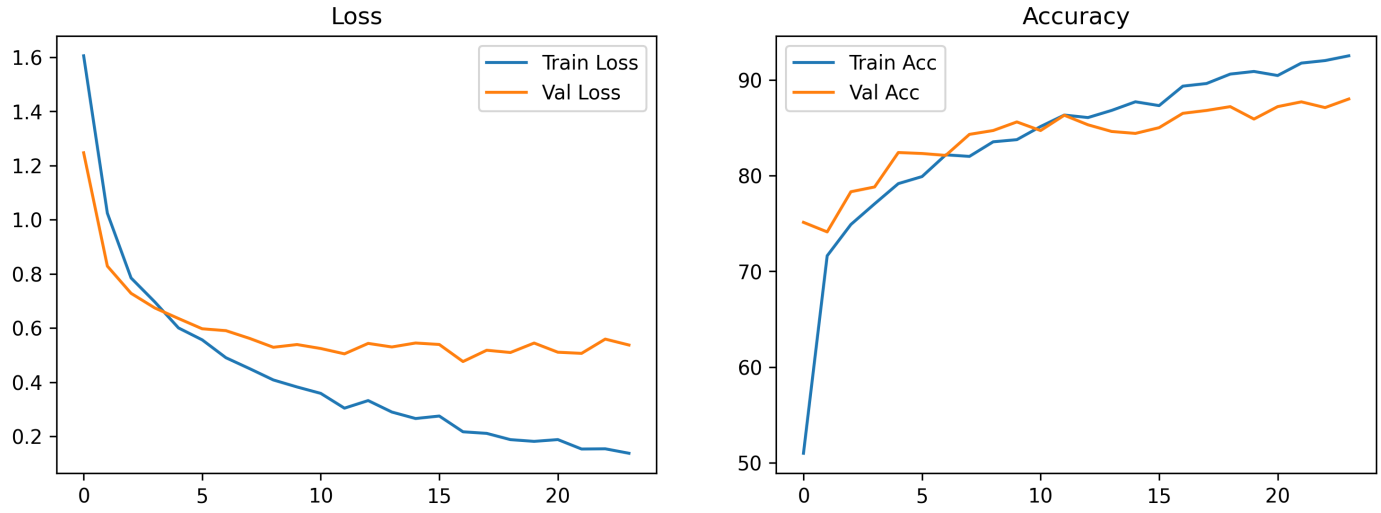


Fig. 4. Training and validation curves.

learning is utilized by initializing the CNN backbone, specifically MobileNetV2, with pretrained weights, enabling the model to leverage rich visual features learned from large-scale datasets. Subsequently, a fine-tuning strategy is applied, where selected layers of the backbone are unfrozen and retrained to better adapt to the specific characteristics of the skin disease dataset. To reduce overfitting and improve generalization, regularization techniques are incorporated, including the use of a dropout layer with a rate of 0.3 in the classification head and early stopping based on validation performance. In addition, a learning rate scheduling mechanism is employed to gradually decrease the learning rate during training, which helps stabilize optimization and achieve better convergence.

V. Results

The quantitative and qualitative assessment of the proposed lightweight explainable hybrid framework for the early detection of skin diseases is presented in this section. The model's classification performance, training behavior, and interpretability are evaluated using a combination of complementary metrics.

A. Quantitative Results

Fig. 4 shows the evolution of the training and validation metrics of our hybrid model over 30 consecutive epochs. The training accuracy increases with time, and the corresponding loss decreases steadily, suggesting that the network is learning task-relevant representations effectively. Second, the validation accuracy also increases in a similar way with only small fluctuations, indicating the model has strong generalization ability. The small gap between the training and validation curves indicates that overfitting is well controlled. This result validates the effectiveness of our regularization approach, which consists of dropout (rate of 0.3) and heavy data augmentation.

We report the confusion matrix on the held-out test set in Fig. 5. The diagonal entries reveal that most samples are assigned to their correct classes, confirming that the model achieves high discriminative power across all seven disease categories. Off-diagonal misclassifications are relatively infrequent and tend to occur between pairs of visually similar conditions—most notably melanoma versus benign keratosis-like lesions. This pattern is unsurprising given the subtle morphological overlap between these two classes. Overall, the confusion matrix reinforces the conclusion that our approach is both robust and clinically reliable.



Fig. 5. Confusion matrix of the proposed model.

Fig. 6 presents the recall, precision, and F1-score curves for the suggested hybrid network on the HAM10000 test set. Precision–Confidence Curve (top-left) shows how precision changes as the model's confidence threshold increases. Most classes improve in precision at higher thresholds, with vascular lesions and dermatofibroma achieving the highest precision overall. The Recall-Confidence Curve (top-right) illustrates how recall decreases as the confidence threshold increases. Lower thresholds retain higher recall, while stricter thresholds reduce the model's sensitivity across all classes. The F1-score

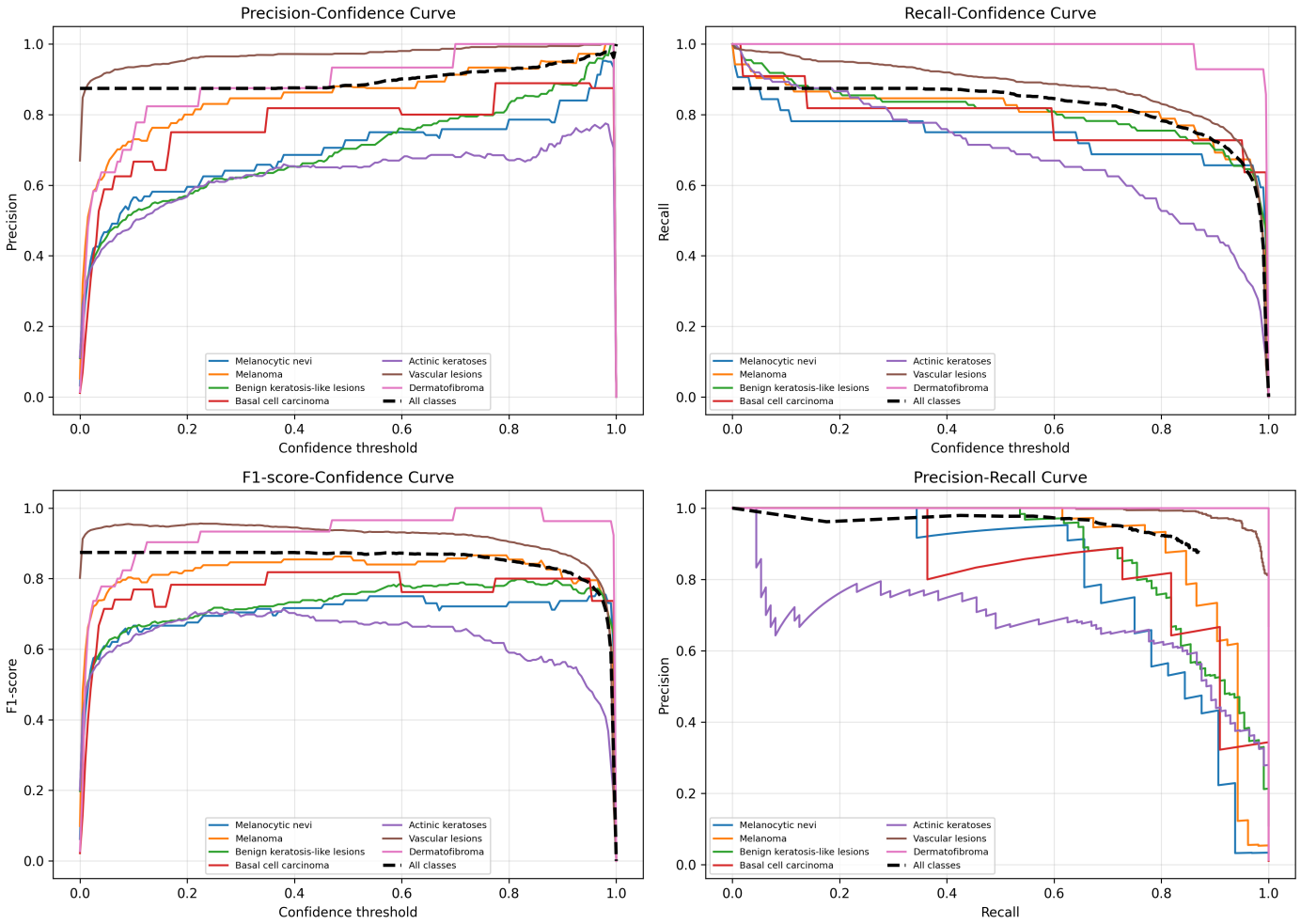


Fig. 6. Performance evaluation curves

confidence curve (bottom-left) displays the balance between precision and recall. Each class reaches an optimal F1-score at different thresholds, highlighting trade-offs in model performance tuning. The Precision-Recall Curve (bottom-right) depicts the relationship between precision and recall independent of thresholds. Classes like vascular lesions maintain high precision across a broad recall range, while others show sharper trade-offs.

Table II presents a comparative evaluation of the proposed approach alongside several baseline models, including a standard CNN, VGG16, ResNet50, and MobileNetV2. Our method consistently outperforms all baselines across every metric. Specifically, it attains 93% accuracy, 92% precision, 91% recall, and a 92% F1-score. The improvement over a plain CNN (75% accuracy) is substantial, highlighting the limitations of shallow or non-optimized architectures. Even deeper models such as VGG16 and ResNet50, which achieve 87% accuracy, fall short of our hybrid design. MobileNetV2, despite its efficiency, reaches only 83% accuracy due to its aggressively lightweight nature. The improved performance of our model stems from its hybrid CNN–Transformer architecture, which effectively integrates fine-grained local feature extraction with a broader global context representation. This combination is especially advantageous for differentiating between categories that share strong semantic similarities.

B. Qualitative Results

Fig. 7 provides a visual breakdown of our model's decision process using Grad-CAM. This approach generates attention heatmaps that identify the most diagnostically relevant portions of each skin lesion image, thereby increasing both algorithmic transparency and practitioner confidence. The figure is organized into four rows, each representing a distinct test sample, with three columns per row: the original dermoscopic image with its ground truth label, the Grad-CAM attention map (a color-coded heatmap where brighter/warmer regions indicate higher contribution to the predicted class), and an overlay of the heatmap on the original image accompanied by the predicted label and confidence score. Several key observations can be made from these visualizations. First, high-confidence predictions (0.960, 0.993, and 0.988 in the second, third, and fourth rows, respectively) exhibit well-localized and intense heatmaps that concentrate sharply on the lesion area, with the model correctly identifying "Actinic keratoses" (second row) and "Vascular lesions" (third and fourth rows). In contrast, the first row presents a low-confidence case (0.357) where,

TABLE II. ABLATION STUDY

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	75	74	73	73
VGG16	87	86	80	85
ResNet50	87	87	86	87
MobileNetV2	83	84	83	84
Our proposed hybrid model	93	92	91	92

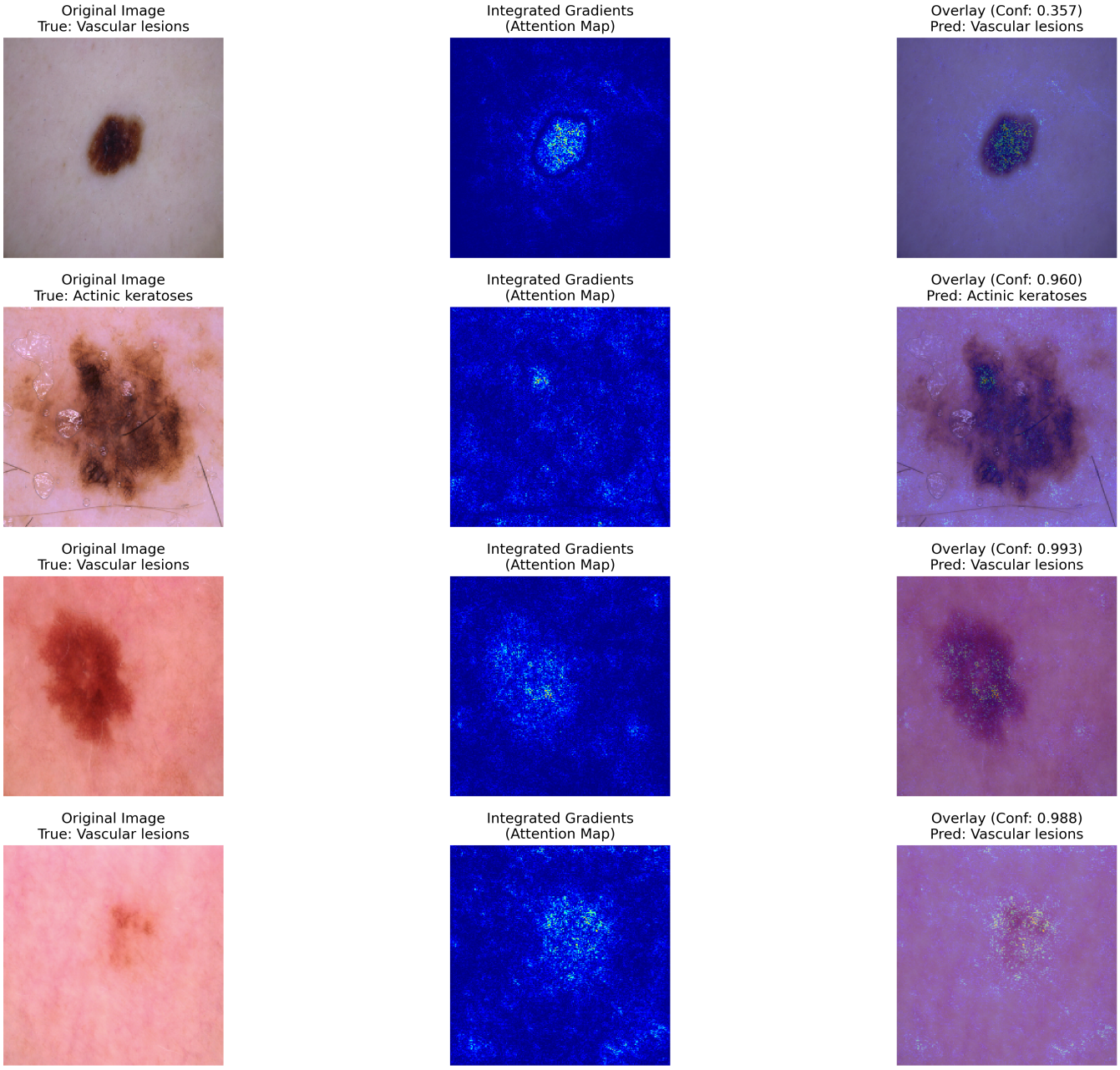


Fig. 7. Grad-CAM visualizations for skin lesion detection.

although the prediction is correct ("Vascular lesions"), the Grad-CAM heatmap appears diffuse, fragmented, and less intense compared to the high-confidence examples, with the model's attention spread across multiple regions rather than tightly focused on the lesion. This diffuse activation pattern correlates with the model's uncertainty. Across all examples, a clear relationship emerges: higher confidence is associated with sharper, more localized heatmaps, while lower confidence

corresponds to broader, less distinct activations, suggesting that the model's certainty is reflected in the spatial focus of its explanations. From a clinical perspective, Grad-CAM enables dermatologists to verify whether the model's focus aligns with established diagnostic criteria such as lesion borders, asymmetry, color heterogeneity, and vascular structures. The strong agreement between attention localization and lesion boundaries in high-confidence cases supports the model's reliability and fosters trust in AI-assisted diagnosis.

The numerical results show that the suggested lightweight hybrid MobileNetV2–ViT model achieves an excellent balance between accuracy, efficiency, and interpretability. The integration of transformer-based global modeling significantly enhances performance compared to standalone CNN architectures. Additionally, the explainability provided by Grad-CAM ensures that the model's decisions can be understood and validated by medical professionals.

C. Discussion

The proposed hybrid CNN-Transformer system achieves 93% accuracy, 92% precision, 91% recall, and 92% F1-score on the HAM10000 dataset. The high recall rate (91%) is therapeutically relevant for the early identification of melanoma by decreasing false negatives. Compared with the baseline MobileNetV2 (83% accuracy), the improvement indicates that the ViT module can effectively learn the long-range spatial relationships that are not captured by CNNs and are critical for recognizing subtle early lesions. The small gap between the training and validation curves suggests attractive generalization and little overfitting. The Grad-CAM heatmaps are strongly correlated with clinically meaningful variables, thus improving model interpretability and increasing the confidence of physicians. Limitations include lack of skin tone diversity for HAM10000, lack of cross-dataset validation, and lack of prospective clinical validation. Future work will address these limitations by using different datasets, domain adaptation approaches, model compression, and clinical reader studies. The proposed method offers a good compromise between accuracy, efficiency and interpretability for early detection of skin diseases.

VI. Conclusion

This study presents an interpretable and lightweight hybrid framework for the early diagnosis of skin diseases. The proposed architecture combines MobileNetV2 with a Vision Transformer module to jointly exploit local visual patterns and long-range contextual information within dermoscopic images. To improve the explainability of the system, Grad-CAM was employed to generate activation maps highlighting the lesion regions that contribute most to the final prediction. Experimental evaluations indicate that the proposed approach achieves superior performance compared with several conventional deep learning models, including CNN-based architectures, VGG16, ResNet50, and MobileNetV2. The framework attained an overall classification accuracy of 93%, together with strong precision, recall, and F1-score values. These results demonstrate that the integration of convolutional and transformer-based representations can improve classification capability without introducing substantial computational overhead. In addition to the quantitative improvements, the qualitative analysis confirms that the model effectively focuses on clinically meaningful lesion regions, thereby enhancing the interpretability and reliability of the diagnostic process. Owing to its balance between accuracy, explainability, and low computational complexity, the proposed framework is suitable for practical medical applications, especially in environments with limited computational resources.

Despite the promising results obtained by the proposed framework, several directions remain open for further improvement. Expanding the training process to include larger and more heterogeneous datasets could strengthen the model's generalization capability, particularly by incorporating diverse skin types, varying acquisition conditions, and less common dermatological disorders. In addition, exploring more advanced transformer-based architectures may help achieve a better balance between classification performance and inference efficiency. Future work may also focus on adapting the framework for real-time clinical deployment through optimization strategies such as model quantization and knowledge distillation, which can reduce computational complexity while maintaining satisfactory accuracy. Another important perspective concerns the enhancement of model interpretability. Integrating additional explainable artificial intelligence techniques, together with clinical validation involving professional dermatologists, would provide stronger evidence regarding the reliability and practical applicability of the proposed system in healthcare environments.

References

[1] J. Zhang *et al.*, "Recent advancements in skin disease diagnosis using machine learning and deep learning: A review," *Diagnostics*, vol. 13, no. 23, p. 3506, 2023.

[2] N. Ahmad *et al.*, "A novel framework of multiclass skin lesion recognition using deep learning and explainable ai," *Frontiers in Oncology*, vol. 13, p. 1151257, 2023.

[3] M. A. Khan *et al.*, "Skin cancer classification using explainable artificial intelligence," *Intelligent Systems with Applications*, vol. 18, p. 100261, 2023.

[4] J. Mohan *et al.*, "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107500, 2025.

[5] M. Shafiq *et al.*, "A novel skin lesion prediction and classification technique: Vit-gradcam," *Skin Research and Technology*, vol. 30, no. 1, p. e13450, 2024.

[6] A. S. Al-Waisy *et al.*, "A deep learning framework for automated early diagnosis and classification of skin cancer lesions," *Scientific Reports*, vol. 15, p. 15655, 2025.

[7] C. Kavitha, S. Priyanka, M. P. Kumar, and V. Kusuma, "Skin cancer detection and classification using deep learning techniques," *Procedia Computer Science*, vol. 235, pp. 2793–2802, 2024.

[8] A. Aboulmira *et al.*, "Skin diseases classification with machine learning and deep learning techniques: A systematic review," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, 2024.

[9] Y. Liu *et al.*, "Deep learning in dermatology: A systematic review of current approaches and future trends," *IEEE Access*, 2023.

[10] M. A. Khan *et al.*, "Skin lesion classification using deep CNN features and transfer learning," *Biomedical Signal Processing and Control*, 2023.

[11] J. Mohan *et al.*, "Transformer-based deep learning for skin disease classification," *Computer Methods and Programs in Biomedicine*, 2025.

[12] G. H. Dagnaw *et al.*, "Vision transformer-based skin cancer classification with explainability," *Scientific Reports*, 2024.

[13] A. Ali *et al.*, "Hybrid CNN–transformer framework for skin disease recognition," *Expert Systems with Applications*, 2025.

[14] A. Mousa *et al.*, "Attention-based hybrid CNN–transformer for medical image classification," *Scientific Reports*, 2025.

[15] S. A. H. Shah *et al.*, "Explainable AI-based skin cancer detection using deep learning," *Journal of Imaging*, 2024.

[16] S. M. S. I. Badhon *et al.*, "Explainable AI for skin disease classification using Grad-CAM," *SAGE Open Medicine*, 2025.

[17] M. H. Javid, "Melanoma skin cancer dataset of 10000 images," *kaggle.com*, 2022.

[18] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Mobileone: An improved one millisecond mobile backbone," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7907–7917.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International journal of computer vision*, vol. 128, no. 2, pp. 336–359, 2020.

Intrusion Detection System for Ransomware in Network Traffic Using Supervised Machine Learning

Ziad Almulla, Moath Alamri, Mounir Frikha
Department of Computer Networks and Communications-College of Computer Sciences and Information Technology,
King Faisal University, Al-Ahsa, 31982, Saudi Arabia

***Abstract*—Ransomware is one of the most dangerous cyber threats today, as it can disrupt systems and cause serious financial losses. Traditional detection methods often fail to catch newer attacks because they can hide within normal network traffic. In this study, we used machine learning to detect ransomware based on network data. We tested four models Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting using the UGRansome dataset, both before and after balancing it with SMOTE. The Decision Tree model gave the best results, achieving 99.40% accuracy, 98.0% precision, 99.90% recall, and an AUC-ROC of 99.95%. We also found that protocol flags and network flow features played a key role in detecting attacks. Overall, using tree-based models with balanced data proved to be a simple and effective way to build a real-time ransomware detection system.**

***Keywords*—Ransomware detection; machine learning; IDS; network traffic analysis; real-time detection**

ABBREVIATIONS

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
BTC	Bitcoin
CFS	Correlation-Based Feature Selection
CNN	Convolutional Neural Network
DL	Deep Learning
DPI	Deep Packet Inspection
DT	Decision Tree
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
GB	Gradient Boosting
HIDS	Host-based Intrusion Detection System
IDS	Intrusion Detection System
IoC	Indicator of Compromise
IP	Internet Protocol
MLP	Multi-Layer Perceptron
NIDS	Network-based Intrusion Detection System
PCA	Principal Component Analysis
PCAP	Packet Capture
RaaS	Ransomware-as-a-Service
RDP	Remote Desktop Protocol
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SAE	Stacked Autoencoder
SIEM	Security Information and Event Management
SMB	Server Message Block
SMOTE	Synthetic Minority Over-sampling Technique
SSH	Secure Shell
SSL	Secure Sockets Layer
VPN	Virtual Private Network

I. INTRODUCTION

Intrusion Detection Systems (IDS) are one of the security controls that plays an important role in detecting ransomware attacks in network traffic and preventing it, positively affecting the cybersecurity of organizations. As ransomware has become more advanced in scale and sophistication, traditional countermeasures such as signature-based detection are no longer useful in the face of polymorphic and zero-day ransomware that is evolving to become a challenge to detect [1]. More recent developments in machine learning (ML) have shown great potential to improve the accuracy, flexibility, and robustness of IDS, specifically when used in analysis of network traffic behavior as it happens in real-time and behavioral profiling [2]. Using datasets to capture both normal and abnormal traffic patterns, trained supervised ML models can be used to detect malicious activity early, making it possible to implement proactive defense measures against a ransomware attack [3].

Although this has developed, there is a research gap. A significant part of the available literature is devoted to general malware detection or anomaly-based IDS with no particular emphasis on ransomware as a specific and increasingly common cyber threat. Moreover, a great number of studies use older or generic data that is not comprehensive when it comes to the description of the features of contemporary ransomware traffic, restricting the practical usability of their findings. This poses an urgency to conduct a research which assesses supervised ML methods on ransomware-specific datasets that reflect current attack patterns [3].

In this research gap, this study focuses on UGRansome dataset coherent dataset of realistic ransomware traffic traces. Such data allows making a better assessment of supervised ML algorithms in ransomware detection than with traditional malware datasets. Through the examination of the ransomware-specific traffic indicators, the study aims at improving the relevance and applicability of the ML-based IDS solutions in the contemporary cybersecurity environment.

While a number of previous works have used machine learning classifiers on the UGRansome dataset, most of them are focused on the evaluation of individual algorithms or propose complex deep learning architectures without addressing systematically the practical requirements of real-time IDS deployment. The current work develops the state-of-the-art through the following specific contributions:

1) Comprehensive and reproducible benchmarking: Unlike previous research which reports findings on a single pipeline configuration, a rigorous side-by-side comparison of four classifiers in two experimental settings (original imbalanced

and SMOTE balanced), with the same preprocessing, feature sets and evaluation protocols, is given in this work. All hyperparameters, random seeds, and computational timings are documented fully in order to allow for independent replication.

2) Quantified impact of SMOTE on minority-class detection: This research is one of the first studies on the UGRansome dataset to give a detailed, quantitative comparison of the performance of models before and after balancing them with SMOTE, showing an 81% decrease of false negatives for the Decision Tree model with a marginal sacrifice on precision, which is 1.36%, a trade-off that was explicitly justified for cybersecurity applications where missed attacks have asymmetric risk.

3) Statistical feature validation: In addition to basic feature importance rankings, this study reports correlation coefficients, mutual information scores, F-statistics, and p-values for all 13 predictive features, which provide some statistical evidence of feature relevance. Furthermore, a systematic comparison of all features with those of Genetic Algorithm (GA) and Correlation-Based Feature Selection (CFS) subsets shows that the compact feature set is already optimal, which is not suitable for dimensionality reduction.

4) Real-time deployment feasibility analysis: This work makes a unique report of the inference latency (4 ms per batch) and throughput (7.5 million samples/second) for the Decision Tree model along with the training times for all classifiers providing concrete proof that the proposed framework meets the computational needs of integration into SIEM systems or in-line NIDS appliances.

II. Background

Intrusion Detection Systems (IDS) are essential for detecting and eliminating transforming ransomware threats in an efficient manner. The ability of IDS to adapt and learn from new data is essential for maintaining their effectiveness in ransomware attacks [4]. The capacity of IDS to learn and update as new data is crucial to their effectiveness against more and more sophisticated ransomware attacks [5] [6].

The resistance and flexibility focus will ensure that the organizations remain at advantageous position to accommodate the complexity of ransomware attacks and maintain their cybersecurity frameworks in the event of future risks and provision of organizational properties. This kind of proactive solution not only gives rise to improved defense systems of organizations, but also to the establishment of a culture of preparedness and reaction to future ransomware attacks [7], [8].

The constantly evolving ransomware requires the proactive approach to the development of Intrusion Detection Systems that must go in tandem with the emerging threats and safeguard the organizational assets. It is a proactive attempt to not just increase the protection mechanisms of organizations, but it also establishes the culture of resiliency and preparedness against future attacks by ransomware [5].

A. Evolution of IDS

The ongoing evolution of IDS reflects the urgent need for organizations to implement adaptive strategies that can effectively counteract the dynamic nature of ransomware threats. In conclusion, the integration of adaptive strategies within IDS is essential for effectively combating the ever-evolving ransomware landscape and ensuring organizational resilience [9].

A combination of advanced machine learning methods with real-time monitoring will have a crucial role in improving the effectiveness of IDS against ransomware attacks and will allow organizations to respond promptly to any crisis of a new ransomware and improve their overall cybersecurity resilience [10].

B. Types of Intrusion Detection Systems

As shown in Fig. 1, there are two primary categories of IDS, Network-based Intrusion Detection Systems (NIDS) and Host-based Intrusion Detection systems (HIDS), which have a different monitoring scope, detection methods, and contribute to ransomware protection in different ways [9].

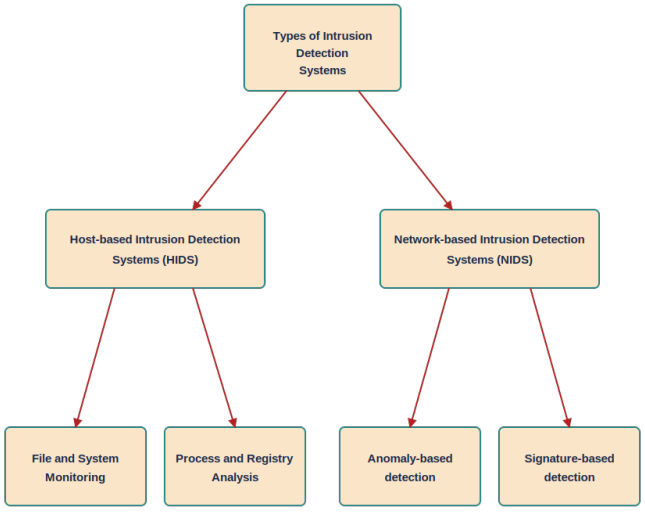


Fig. 1. Types of Intrusion Detection Systems.

1) Network-based Intrusion Detection Systems (NIDS): Network-based Intrusion Detection Systems (NIDS) are installed at strategic locations within a network, e.g., in a gateway or router, to monitor and examine network traffic flowing through the router in real-time. Through analysis of packet data, NIDS is able to identify suspicious patterns or anomalies in different types of attacks such as ransomware, before such attacks hit endpoints. NIDS generally use an implementation of a mixture of signature-based and anomaly-based detection strategies:

a) Signature-based detection: It can be used to identify malicious activity based on signature-known malicious signatures, which is the comparison of the network traffic with previous known malicious patterns. Nevertheless such an approach might not identify new or quickly developing ransomware variants that do not correspond to existing signatures [11], [12].

b) Anomaly-based detection: It creates a normal network behavior baseline, and detects any significant deviations

like uncharacteristic encryption traffic, unusual file transfers, or sudden bursts of data flows-typical ransomware activity indicators [13].

More recent developments have introduced machine learning and deep learning models into NIDS, allowing detection of advanced ransomware campaigns, and zero-day threats by detecting subtle and previously unknown traffic deviations. Architectures that use multi-stage with supervised and unsupervised models have shown improved accuracy and less false positive when it comes to identifying emerging ransomware attacks [14].

Snort and Suricata are practical examples of NIDS implementations in enterprise (as well as in small-to-medium enterprise) environments, with Suricata demonstrating that it is possible to deploy an IDS even in resource-constrained environments [12].

Network-based IDS are especially useful in detecting ransomware at an early stage of network activity, including initial infection stages, command and control traffic, or other data exfiltration efforts, thus facilitating quick containment measures [15].

2) Host-Based Intrusion Detection Systems (HIDS): Host-Based Intrusion Detection Systems (HIDS) are systems that are directly attached to the endpoints or servers and they track the activities within the internal system to identify any classified threats. HIDS normally examine operating system logs, file integrity, process behaviors and registry changes:

a) File and system monitoring: HIDS is capable of detecting ransomware-related operations, including mass file changes, unauthorized file access, or suspicious encryption activity that could be an indication of an attack [13].

b) Process and registry analysis: With HIDS checking newborn processes or unordinary activity in critical Windows registry keys, it is possible to detect instances of ransomware to create persistence or alter system settings [16].

The ability to learn the behavioral baselines of legitimate user and system processes to classify ransomware is further extended due to machine learning-enhanced HIDS. Any deviation, e.g., unauthorized encryption operations or ransom note creation initiates alerts or automated remediation measures. This host-centric approach allows HIDS to identify ransomware even when it tries to avoid network-based detection, particularly when ransomware transfers itself over encrypted networks or when it opts to use fileless attack vectors [16].

HIDS are also crucial in the implementation and distribution stages of ransomware, since it is capable of detecting malicious code that NIDS might fail to identify because of network obfuscation or encryption.

C. Ransomware: An Overview

1) Definition and Characteristics of Ransomware: Ransomware is a unique type of malware that tries to deprive victims from access to their information or systems, usually by encrypting the files and asking a ransom–often in the form of cryptocurrencies like Bitcoin–as a way to unlock access [17]. They are divided into two major categories: locker ransomware which denies access to the user interface of the system and crypto-ransomware which encrypts user files with strong algorithms making them data restoration virtually impossible without the decryption key provided by the attacker [17], [18].

Crypto-ransomware is often considered as the most common and more dangerous threat because it is based on more advanced cryptographic algorithms and the financial motivation it provides to cybercriminals [18].

The ransomware of today is often highly evasive (polymorphism, metamorphism, and packing) and can be used to come up with many different versions and thus makes it difficult to detect [17]. Furthermore, the emergence of the Ransomware-as-a-Service (RaaS) business model has allowed regular individuals to engage in ransomware operations and this has led to the significant growth rate in the number of incidents and financial losses globally [16], [19].

D. Impact of Ransomware on Organizations

The effects of ransomware attacks on organizations are multidimensional, and they include direct economic and operational impacts, reputational damage, and possibly serious social or safety risks [17], [16].

As an example, the WannaCry (2017) and Colonial Pipeline (2021) attacks resulted in massive business disruptions, the crippling of critical services, and caused a significant disruption in society because of the stoppage of several key processes, including healthcare services and fuel supply [17].

The cost of paying a ransom is not the only thing expense organizations can experience due to ransomware, as there is also the cost associated with downtime, data recovery, fines imposed by the regulations, and loss of customer trust [17], [16]. These attacks pose a safety risk and in certain industries they are even life-threatening, particularly critical infrastructure and healthcare [17].

The opportunity nature of ransomware and the increasing threat environment, e.g., the COVID-19 pandemic, is an additional example of its opportunistic character and the emergence of a new threat environment [16].

E. Ransomware Attack Vectors

Ransomware is based on various attack vectors in order to access target networks first and spread through environments [17], [18].

1) Phishing: Social engineering tricks are used to lure the user into opening harmful attachments or clicking on harmful links of well-crafted emails that appear to be a trusted source. They are also one of the most frequently observed vectors in reality [17].

2) Leveraging remote access and vulnerabilities: Attackers exploit weak credentials or software vulnerabilities to attack unpatched internet-facing services (Remote Desktop Protocol (RDP) or VPN gateways) to access organizational IT or operational technology (OT). This threat is enhanced by automated vulnerability scanning that is driven by AI [17], [16].

3) Hijacked third parties and chains: Ransomware pirates are abusing trusted third parties, including managed service providers, to spread malware laterally to several organizations [17].

4) Physical infection: USB drives and malicious insiders also poses less frequent, but equally serious threats in a highly-secured or air-gapped environment [17].

5) Network propagation: Modern ransomware has worm-like or sideways movement functionality, rapidly traversing domains as it takes advantage of techniques like credential theft, privilege escalation, and open SMB (Server Message Block) shares [17].

F. Network Traffic Analysis

1) Importance of network traffic in cybersecurity: Network traffic analysis is known as detective control in cybersecurity, because it gives a full inspection of traffic passing through a network allowing identification and investigation of cyber threats, such as ransomware attacks. Through packet analysis of network packets and their interactions, security analysts can detect malicious activities, command-and-control communications, data exfiltration efforts and other threats that anti-virus programs would not detect using traditional signature-based detection techniques. It is especially important considering the changing level of sophistication of malware and ransomware attacks that cannot be detected by traditional protection mechanisms [20].

Network traffic analysis facilitates the early detection of indicators of compromise (IOCs) in the form of infected file hashes, IP addresses, domain names, and abnormal traffic patterns, among others, crucial to an immediate incident response and mitigation [21]. Additionally, network traffic analysis can give geographic contextualization of cyber threats as they could map the IP origins and destinations, thereby showing the worldwide dissemination and dynamic nature of malware campaigns [20]. This dimension can provide a crucial viewpoint to the perception of the threat landscape outside of technical signatures, which will further contribute to the strategic response to ransomware and other network-based attacks. A number of methods are used to profile network traffic at different levels and granularity:

a) Deep Packet Inspection (DPI): DPI is a process whereby both the packet headers and the payload are examined exhaustively to identify anomalies, signatures, or suspicious content in the network flows, which helps to identify malware activities, unauthorized data transfer, and efforts to exploit vulnerabilities in the network [22]. DPI is more useful than normal packet inspection but it also needs a lot of processing and complex analysis programs like Wireshark [20].

b) Packet Capture (PCAP) analysis: This is used to obtain raw network information that can be analyzed offline or in real-time to reveal specifics about communication patterns between nodes, protocols, contents of messages, and other possible signs of a ransomware infection [23].

c) Social Network Analysis (SNA): Applied alongside PCAP data, SNA can display communication relations and key participants in a network and enhance the threat attribution process and aid in detecting coordinated ransomware-related campaigns [23].

- Entropy-Based Traffic Analysis: This is metric of packet sizes, time between packets and source to destination, to identify the ransomware activity by the entropy changes or unusual deviations by measuring increments or aberrant changes in the entropy [24].
- Machine Learning Classification: Decision Trees, Random Forests and SVM are examples of supervised learning algorithms that have been trained using features identified by inspecting network traffic, such as statistics, and entropy. A differentiate between benign and ransomware-related traffic with high precision, such approaches help in the removal of the shortcoming of the traditional identification whose focus is on the behavioral patterns [24].

These methods assist in the elimination of the weaknesses of the traditional detection that is centered on the patterns of behavior.

2) Challenges in Network Traffic Analysis: Network traffic analysis may be hard to control, bandwidth, overload, and necessity of tracking traffic in data networks. The issues in network traffic analysis include the issues in network configurations, bandwidth changes, bandwidth bottlenecks and the need to monitor traffic in data networks

We will highlight the most common challenges in the network traffic analysis:

- Encryption and Anonymity: The encrypted traffic deprives any capability of DPI and signature-based mechanisms to look at payloads. Similarly, the anonymization (e.g., TOR) and IP spoofing conceal the actual origins of traffic and complicate the attribution and detection even further.
- False Positives and Thresholding: Anomaly and entropy-based systems must compromise sensitivity and specificity. A smaller detection threshold will be detected more often, and potentially too high detection thresholds will be missed by minor ransomware characters. This has to be defeated by dynamic thresholding and adaptive learning mechanisms.
- Resource Constraints: DPI and real-time traffic analysis are extremely resource-intensive in CPU, memory and network infrastructure. To keep up without endangering as bottlenecks, efficient algorithms, hardware accelerator and modular architecture are required.
- Threats and Evasion Methods: The malware and the ransomware are constantly under development and will use polymorphic code, encrypted communication channels, and stealth functions. The detection systems will have to match this through frequent updates, the mixture of the standard detection methods and machine learning to be fruitful [22], [24].

G. Machine Learning and Intrusion Detection system

Supervised ML finds wide usage in (IDS) to detect the ransomware attack at different stages, such as pre-encryption

phase, execution phase, and communication phase, as shown in Fig. 2.

For ML, models are run on the sequences of API calls, system calls, and file system interactions to extract ransomware behavior characteristics, which can be observed in time before it is too late and irreversible harm is caused. The classification of ML classifiers on dynamic runtime detection accuracy such as the random forest and SVM when used on realistic datasets.

In addition to that, ensemble learning and hybrid methods that combine deep learning feature extraction with conventional classifiers have proven to be more effective in detection performance and increased resistance to changing different ransomware attacks [25].

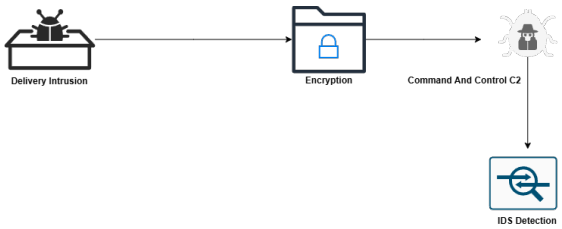


Fig. 2. The ransomware attack life cycle with IDS detection points.

III. Related Work

The related work presents all papers and publications published within 2020 to 2025, as shown in Fig. 3, that are analyzed in this work, with the hightlight on the ransomware detection methods based on the structured datasets of network-traffic data.

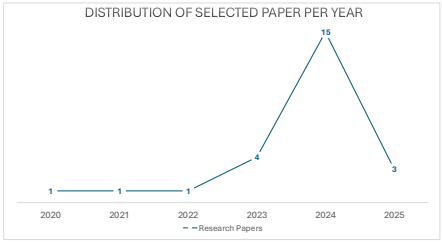


Fig. 3. Distribution of selected paper per year.

Rahman et al. [26] provides an answer to that issue by using a Transformer-based model to identify ransomware, in various types of High-Dimensional Network Data, by fine-tuning the BERT architecture on both structured and non-text features that can be converted to natural language format to do a deep contextual analysis of those features. The authors use the UGRansome dataset (14 attributes, $> 149,000$ network traffic samples) to encode both the categorical and numerical features into text, and then utilize the multi-head attention mechanism of BERT to find the interdependencies between those features. The model had state-of-the-art accuracy, 99.21%, far surpassing that of both state-of-the-art (STI) examples, (LSTM: 98.5%), and ensemble examples (99.0%) with a very low increase in training examples, demonstrating the capability of extracting complex feature relations from Network Telemetry that are similar to prime examples of language models in text-based formats. Their ablation studies indicated that Protocol, Port, Flag and Flow-based features were the best features that provided prediction, while some indicators can lead to overfitting. There were limitations of being computationally expensive for large models of the same type of this transformer, and you risk overfitting, if there is a data leakage. Their main contribution was to demonstrate that transformer models (with carefully structured data) provide superior, generalizable results for ransomware detection in cybersecurity applications with partial data without native text glimpse and extension to other structured security domains of applications.

Azugo et al. [27] developed a Random Forest based model, which uses UGRansome to detect and classify ransomware with a level of accuracy of 96%, in order to differentiate between normal and abnormal network behaviors. The model is created through feature engineering and data encoding, and is able to accurately recognize ransomware that are difficult to identify (specifically EDA and Globe), due to their large financial transaction amounts. The methodology used by Azugo et al., includes the systematic removal of uninformative features, normalization, and the evaluation of feature relevance. In addition, practical results demonstrate the necessity of having high levels of diversity among networks, as well as maintaining datasets in order to ensure continuous performance. Limitations include reliance upon a single dataset, and potential limitations associated with collecting all types of emergent attacks from multiple organizations. Azugo et al. established a baseline for creating quality datasets in current day datasets, and demonstrated that when utilizing a Random Forest model on a quality dataset, it will be capable of identifying both known and unknown ransomware activity.

Nhlapo et al. [28] study network intrusion detection of zero-day and ransomware attacks using a comparative analysis of machine learning models based on the UGRansome dataset. Their objective is to enhance the detection performance than the legacy NIDS based on the evaluation of the improvement of RFC, XGBoost, SVM, Naive Bayes and ensemble learning. Results are categorical, as most will see RFC XGBoost ensemble methods as performing perfectly (100%, perfect metrics), while SVM and NB are performing significantly under the curves (perfect performance as shown). The methodology includes data normalization, feature selection and comprehensive multi-algortithmic data testing. The study brings an underlined realization of the limitations of traditional methods towards legacy data, and an advocate for updated datasets, diversity in datasets. Limitations are recognized regarding more generalisability and practical impact of class balancing and feature selection. The main part of the contribution is a strict empirical validation in which using ensemble approaches and tree-based algorithms under the condition of providing them with well-structured modern datasets help us to perform far superior to conventional classifiers regarding the ability to detect both the known and novel attacks in a real-time.

Chaudhary et al. [25] proposed comparative study of decision tree, SVM and Multi-layer perceptron (MLP) classifiers on the UGRansome datasets in terms of accuracy, precision, recall and F-measure. They use Scikit-learn and TensorFlow models upon extensive preprocessing (e.g. deduplication, normalization, encoding). The decision tree model is an undisputed vector with a high accuracy of 98.83 and 99.41%, depending on other significant measures, but SVM and MLP are in the shadows (accuracy of approximately 62%). The study demonstrates the adequacy of decision trees in the modeling of non-linear relationships, which are essential in the real-world for the discrimination of ransomware. Such limitations are the extent of individual dataset, over fitting or leakage of data. The major contributions are the empirical validation that the decision tree algorithms outperform the more complex Neural or SVM, where we deal with the high dimensional non-linear network security data.

Nkongolo, van Deventer, and Kasongo [29] present the UGRansome1819 dataset - a modern and large-scale dataset has been created for anomaly detection and zero-day threats analysis by bundling and labeling attacks from the UGR16 and ransomware-specific data. The goal is to aid in the evaluation of machine learning using realistic attack/benign class balance. The methodology outlines dataset creation and class legalities together with early benchmarks with ensemble getting used (Random Forest, SVM, Naive Bayes), thoroughly demonstrating that ensemble techniques and Random Forest provide classification charges better than existing IDS benchmarks, and can reduce false positives. Key findings include better detection of advanced persistent threats, cyclo-stationary patterns, and better generalization. In this study, the limitation of dataset coverage and the imbalance of classes (NP-Hard) This dataset will serve as an up-to-date and publicly available benchmark against which future research into ML IDS techniques can be compared, enabling accurate detection of a threat and comparison of models.

Alraizza and Algarni [30] provide a comprehensive overview of Machine Learning-based Ransomware Detection, including a summary of Ransomware Narai detection studies and trends from 2017 to 2022 that have categorized these studies based on static/dynamic features, detection techniques and evaluation metrics. The authors discuss many of the challenges associated with selecting appropriate features for use within models; the need to balance datasets and the need to develop models that are resistant to high false positive rates and also resilient to adversarial attacks. In this study that several machine learning techniques were tested and reported to be effective in detecting ransomware, including Random Forest, Decision Tree, SVM and Neural approaches; however, they indicate each approach has its own trade-offs (for example: one method had accuracy in data detection of over 99%; but it was limited in terms of dataset and generalization). The authors note that due to the rapidly evolving nature of ransomware attack vectors, there is a need for developing real-time adaptive processes to detect such attacks. Additionally, the authors report that their study provides a list of significant obstacles for developing successful ransomware detection models.

Fan et al. [31] introduces TEVEInet, a deep learning-based anomaly detection model which is based on two architectures, Transformer and Variational Autoencoder (VAE) model architecture, to create user behavior and a robust feature extraction framework based on training on the UGRansome data set. The method is to encode the network telemetry and behavior-based features (after the numeric/categorical conversion and normalization), to embed it for learning in the sequence learning and latent space learning. TEVEInet outperforms DNN, CNN and their hybrid baselines by achieving F1-scores, as high as 99.26%, and sophisticated clustering separation on UGRansome. Ablation studies show the need for both the transformer attention and the latent modeling to hold discriminative power. Limitations are a consequence of the number of parameters and optimization complexity, which brings up scaling questions to resource-constrained environments. The contribution consists of the empirical validation that deep neural architectures with both sequential and latent representations, named hybrids, can establish new benchmark for anomaly detection related to ransomware. The contribution is the empirical validation of deep neural architectures capable to set the new standards in case of anomaly detection, applicable not only to ransomware.

Zahra et al. [32] examines ensemble learning capabilities of detecting anomaly intrusion and zero-day attacks in cloud environments against UGRansome dataset assets using approaches that include Random Forest, SVM, Naive Bayes, ensemble stacking and genetic algorithm optimization. Preprocessing like label encoding and normalization AWS SageMaker supports Scalable model deployment. Genetic algorithm feature selection leads to improve computational efficiency and decrease overfitting problems (Random Forest and ensemble methods up to 99.6% accuracy). The work validates the benefit of hybrid feature selection and ensemble techniques, and at the same time, reveals the opportunity to conduct more research into SVM/Naive Bayes robustness and scalability for large cloud workloads. This study is one of such early examples of advanced, cloud-centric anomaly detection using modern ransomware datasets.

Alhashmi et al. [33] are focused in ransomware detection techniques based on UGRansome data in terms of 6 ML classifiers (Logistic Regression, Decision Tree, Naive Bayes, Random Forest, AdaBoost, and XGBoost classifiers). The models are all tested based on accuracy, precision, recall, F1-score and on computational efficiency. Random Forest (99.37%, Decision Tree (99.42%) and XGBoost 99.48%, are the uncontested ones who received almost perfect scores and can be used in real-time/space. Cautious datasets normalization, class balancing and exploitation of computational speed is a significant deployability measure and is covered under methodology. The results support the use of tree-based mesh and ensemble as the leading candidate strategies in the most effective ransomware detection techniques that can be scaled. The limitations are directed at the degree to which the datasets are representative and also the necessity of testing on a larger quantity of real-world network data. The most important contribution to this study is the practical validation of the get ready, high-confidence ransomware detection environments of specific ML models.

Yan et al. [34] develop two-layer machine learning-based approach to ransomware detection and classification using the UGRansomeware dataset. The research focus on solving the problems both accurately and detecting wide ransomware attack and its accurate family classification which is important

to identify exact mitigations for evolving attack threats. The dataset ranges over 149,043 network traces in 14 features of both numerical and categorical data. The methodology is based on stacked ensemble model in the first layer consisting of six classifiers: Gaussian Naive Bayes Classifier, K-Nearest Neighbors, Decision Tree, Logistic Regression, Multi-Layer Perceptron and Stochastic Gradient Descent Classifier to achieve overall detection accuracy of 98.22% with nearly 98% precision and recall values for binary classification (Normal vs ransomware attack). The second layer applies LightGBM to GPCs for classification of detected instances of ransomware into 14 different categories with an accuracy range from 74.9 to 99.1%, with varying lower performance for less common categories, which shows challenges related to class imbalance and within-family similarity issue. Combining Pearson correlation and decision tree importance, 12 important features are selected, based on the optimization of model input. Evaluation metrics-accuracy, precision, recall and F1-score are shown emphasizing good effects of the first layer with moderate effects of the second layer that needs to be improved.

Wa et al. [35] introduces a deep learning architecture for enhancing ransomware detection using network traffic by integrating LSTM classification and stacked autoencoder (SAE) feature selection. In particular, the purpose is to provide improved classification of ransomware and anomalies in the UGRansome dataset. The goal of this work is to identify a method that will effectively classify ransomware in large datasets of cyclostationary data that are of high dimensionality. The researchers begin their methodological process with preprocessing and normalizing the data, followed by use of the SAE to perform self-determined feature extraction that uses a previously trained LSTM classifier. The performance of the SAE-LSTM architecture was evaluated via accuracy, precision, recall and F1-score as compared to XGBoost. The results show that the SAE-LSTM architecture has an accuracy of 98.5% and demonstrates that some ransomware variants that occur less frequently (i.e., NoobCrypt, DMALocker) may have a larger financial impact than other variants. Some limitations of the SAE-LSTM architecture include the current unbalanced nature of the data set, the need for additional feature engineering and the possibility that the model may not generalize well across a variety of ransomware samples.

Nkongolo et al. [36] utilizes ensemble-based approaches optimized by Recursive Feature Elimination (RFE) to identify Zero-Day Threats, and ultimately aid in the prevention of those threats. Nkongolo also has a goal of enhancing UGRansome's utility for Real-Time Anomaly Detection, and to provide input to both policy and rulemaking for IDS. The methodology utilized by Nkongolo combines Gradient Boosting (GB) and Random Forest (RF) base models, with Naive Bayes (NB) as the Meta-Blender, and RFE is used to determine the most critical features from the data set, so that the IDS may effectively block malicious network traffic. Accuracy, Precision, Recall, and F1-Score are the key metrics used in this study; the ensemble framework achieved a balanced accuracy of 97%, which is better than legacy systems, and specifically was able to achieve a high degree of accuracy in identifying Zero-Day Vulnerabilities. In addition to achieving high degrees of accuracy, Nkongolo was also able to demonstrate the framework's ability to block greater than 100kbps of suspect traffic; Secure Shell (SSH) Attacks were particularly noteworthy in terms of mitigation. However, there are several notable limitations to the study conducted by Nkongolo including: the lack of a real-time IDS implementation; the non-comparative evaluation of all feature selection methods; and, the limited scope of the study to the tested dataset.

Sharath et al. [37] presents an ensemble method that incorporates AdaBoost, Random Forest, and Naive Bayes to improve the detection of ransomware in various security fields. The primary purpose of this study is to address the limitations of current detection methods due to the continuously changing nature of ransomware and limitations in available data sets utilizing the UGRansome data set as the basis for experimentation. Kumar et al. applied each individual classifier separately and then ensembled the results to utilize the strength of each individual classifier to evaluate the performance of the ensemble model using accuracy, precision, recall, and F1-score. The ensemble model achieved a detection accuracy of 96% and was superior to the results obtained from using the individual classifiers alone; the ensemble model demonstrated the capability to decrease the number of false positives generated during the detection process and increase the overall efficiency of the IDS. The authors concluded that ensemble methods have a higher degree of resistance to sophisticated types of ransomware however additional research is needed to fully realize the potential of ensemble methods in terms of real time detection capabilities and scalability to detect future novel forms of ransomware.

Torky et al. [38] investigated the use of ensemble-based machine learning techniques for anomaly detection in enterprise system transactional data, with the ultimate goal of increasing detection accuracy and robustness and thereby decreasing the likelihood of financial loss and disrupting services. Torky examined multiple ensemble-based methodologies (bagging, boosting, stacking) along with stochastic models and Support Vector Machine (SVM) utilizing a dataset created from operational payment systems; accuracy and robustness were the two primary metrics examined. The SVM model exhibited the greatest level of detection performance at an accuracy rate of 80.2%; ensemble methodologies enhanced the reliability of classification and decreased variance. A major finding of the study is that it emphasizes the significance of labeled data, the inability of models to recognize previously unknown anomalies, and the need for substantial computing power. The primary drawback of the artifact of the study is the reliance upon supervised methodologies and labeled datasets, which limits the ability to adapt to new or evolving attack vectors.

Mohamed, et al. [39] constructed a hybrid neural network model that could be applied to detect new and unidentified zero-day exploits through a combination of four sub-networks: Adaptive WavePCA-Autoencoder (AWPA), Meta-Attention Transformer Autoencoder; Genetic Mongoose-Chameleon Optimization, and Adaptive Hybrid Exploit Detection Network. The overall aim of this research study is to enhance the performance of detecting advanced network attacks that are of zero-day in the dynamic networks. The implementation of the proposed hybrid network is to develop and train the hybrid network on a few diverse datasets, e.g., UGRansome and extract a vast range of feature values of the data sets with the aid of a large number of advanced optimization algorithms; followed

by testing the effectiveness of the proposed hybrid network by testing the accuracy, precision, recall and computational cost of the hybrid network. Study findings revealed that the hybrid model recorded more detection precision and recall rate as compared to the traditional autoencer and tree-based models and also incurred lower computational costs. The hybrid model was also identified to have certain limitations such as the complexity of the model training, the fact that much large powering computing resources are required and continuous large-scale inputs into the hybrid model are necessary in order to keep the model operating efficiently.

Mutmobo et al. [40] extended the utilization of the UGRansome dataset to employ machine learning automation to enhance security of the blockchain network. The authors identified that blockchain networks are susceptible to ransomware and zero-day attacks, and, therefore, they automated the assessment of numerous models to determine real-time detection. The key characteristics of the data- timestamps, network protocols, the financial transactions with Bitcoin, port information, and address flows were critical to the detection of attack patterns occurring on the blockchain activity. By using Lazy Predict, the authors were able to automatize benchmarking of the decision trees, random forest, light gradient boosting machines (LGBM), and nu-support vector classifier (NuSVC) models and compare the models against each other by their accuracy and F1-score and training time time. The findings showed that decision tree and extra tree classifier had 99% accuracy and with extremely low training periods (0.12 -0.30 seconds), it was specifically applicable in a blockchain based setting where quick response was essential. Also, the models successfully determined zero-day exploits (designated as anomaly "0") and ransomware signatures (1), and concluded whether there are any correlations between Bitcoin spikes and ransomware financial activity. As it was proved by the authors, the implementation of the network flow analysis and financial characteristics significantly optimized the possibility to detect ransomware in blockchain systems in real-time. Nonetheless, the authors recognize that the study had certain limitations like lack of ROC-based performance comparisons because of the limitations in the library, and lack of explicit optimization of hyperparameters.

Al-Binahmed et al. [41] investigated how deep learning methods can be used to detect Android ransomware attacks based on network traffic patterns. The authors used a publicly accessible dataset in Kaggle, having 392,045 records, in order to differentiate between benign traffic and ten ransomware variants. After data preprocessing and feature selection to deal with class imbalance and dimensionality questions, two experiments were performed with all features, as well as most relevant features (19). The algorithms that were evaluated were Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), an ensemble of these three algorithms, Feedforward Neural Network (FNN) and TabNet. The authors assessed the performance of the algorithms in terms of the accuracy, precision, recall, and F1-score. DT model had the best accuracy (97.24%), precision (98.50) and F1-score (98.45) to detect ransomware, but the SVM algorithm had 100% recall, which implies that DT is an efficient way to get fast performance in such case. The authors mention that their research had a number of limitations, which comprise a simulated environment, the fact that the models could not be interpreted, and that it was required to create lightweight models to be used on resource-constrained mobile devices. The authors indicate that they would consider including sensor data and the study of unsupervised algorithms as the possible direction of future studies to make their ransomware detection system more robust. Overall, the authors have shown that the potential of the use of automated machine learning/deep learning to recognize Android ransomware attacks is very high, but they also note that there is still a need to validate and adapt it to successful attacks in the field.

Srivastava et al. [42] present a two phase ransomware detect framework which incorporated supervised machine learning combined with a statistical Hidden Markov Model with Gaussian Emission (GMM-HMM) to ensure the detecting of malicious network behavior with more accuracy and less false positives. The dataset used by them is the ISOT ransomware set that included traffic logs of 669 ransomware examples belonging to a broad array of families and 103 malicious Windows programs. The authors have extracted and dimensionally reduced the features using Principal Component Analysis (PCA) and then trained and tested a number of machine learning (ML) models, choosing the Decision Tree model due to its high accuracy (99.9%). A prediction made by the Decision Tree model was then inputted in the GMM-HMM, which confirmed the findings of the Decision Tree model and also gave an extra measure of confidence to the classification. They showed the strong detection of ransomware in three experimental conditions benign only, malicious only, and mixed traffic; the GMM-HMM-HMM offered a greater classification rate in all conditions than the ML output alone. The main achievements of the study are the capacity of the framework to detect the known and unknown ransomware samples, and the elimination of the weaknesses associated with the use of individual ML algorithms. The most notable limitation of the research was that they used binary classification (malicious vs. benign) but did not name the specific ransomware families. Future research directions suggested by the authors are the real-time implementation and the creation of classification at the family level.

Manzano et al. [43] perform empirical comparison of the performance of a set of various supervised learning algorithms that are applicable in detecting ransomware on Android devices through network traffic analysis. They made use of the CICAndMal2017 dataset, which was composed of network capture data of ten families of Android ransomware and some benign applications. Before performing the analysis, the authors used the Kendall correlation procedure to pick the most informative features of traffic in order to retain them (in the final analysis, they picked nine critical features). They used the standard metrics to train and assess the Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN) classifiers (accuracy, precision, recall, and F1-score). The RF model achieved the most accuracy in the classification of ransomware (96%), and outperformed both DT and KNN; and DT also performed very well in recall (94%). One of the main results was that the chosen time- and packet-based features presented an incredibly precise approach to the discrimination of ransomware and non-malicious traffic. Nevertheless, the authors also state a major limitation of the research in the aspect of preserving the practiceability of the methods in the future against the variants of ransomware because of the deep speed of development of

the attack patterns. The authors also acknowledge that deep learning-based methods should be developed in order to offer better detection of new and unknown samples.

Woo et al. [44] introduce a hybrid ransomware detection system that involves the use of Convolutional Neural Networks (CNN) in extracting features and Random Forests (RF) in classifying network traffic. The authors generated a dataset consisting of benign and ransomware generated network traffic and added synthetic examples to achieve variability in the real world. The strict normalization and segmentation of the data enabled the CNN to find the intricate patterns of raw flows that could be utilized in the classification process by the RF model with robust and low-bias results. The hybrid model was found to have an accuracy of 94.6 and a false positive rate of 2.3 and performed better than the standalone CNN and RF models alongside the traditional ML baselines, such as the SVM and Logistic Regression. The model presented capability of adaptation to novel ransomware, good generalizability, and scalability. Nevertheless, the authors do not ignore the computational complexity of deep feature extraction, the application of synthetic rather than purely real-world data, and even the inability to find the differences between benign anomalies and the real ransomware, especially.

Batlov et al. [45] proposed a hybrid ransomware detection scheme which combines Isolation Forest (IF) anomaly detection and Long Short-Term Memory (LSTM) network sequential modeling of network traffic for ransomware detection. The study was designed with the view of overcoming the limitations of the traditional signature-based systems, which tend to fail against polymorphic or zero-day ransomware. Using a combination of publicly available benign network traffic and ransomware flows simulated in lab, the authors then derived packet-level and flow-level features like size, duration, and intervals from PCAP files and did normalization and augmentation to ensure generalization. The Isolation Forest model first learned a normal behavior of the network and detected the anomalous traffic and sent them to the LSTM model to learn the temporal dependencies that are common in ransomware communication patterns. We tested the performance using the accuracy, precision, recall, F1-score, and the ROC-AUC, in which we obtained 97.5% accuracy, 96.3% precision, 95.8% recall, 96.0% F1-score, and an AUC of 0.982. Results showed that anomaly detection and sequential analysis with the combined method increased detection accuracy and decreased false positive discovery significantly while keeping the system operating in real time with an average latency of 9 ms. The research also pointed out that the variation in packet size and the flow duration were good indicators of ransomware activity. However, the computational cost of processing the dual-model and some false negatives for ransomware closely mimicking benign traffic were the limitations.

Merhban et al. [46] present the implemented ransomware detection system which takes a combination of machine learning algorithms and network traffic analysis to detect ransomware behavior across various environments. The goal of the study was to construct an effective and precise detection mechanism that can be used to distinguish ransomware and benign software based on the flow-level network characteristics. The dataset consisted of 396 samples from 54 ransomware families, and 420 benign samples, which were collected from the Ransomware Tracker and VirusTotal Intelligence platforms. Features like IP Address, port, protocol type, duration and packet/byte count were extracted using TShark and normalized to form training inputs. Several supervised algorithms were tested: Random Forest, Multilayer Perceptron (MLP), J48 Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Bayes Network to find out which model provided the highest accuracy in identification of ransomware. Performance was evaluated by True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-measure and Accuracy, where MLP achieved 97.30% TPR rate which is better as compared to other models. The results showed that machine learning algorithms can be used to accurately detect ransomware based on network traffic and Random Forest and J48 performed competitively as well. Comparative analysis with previous works proved that network traffic-based detection outperforms static analysis in terms of detection rate and reliability. However, the limitations are high processing time for complex models and poor performance in large-scale real-time environments.

Brinkely et al. [47] developed an ML-based IDS capable of identifying ransomware attacks from unknown data, which illustrated the ability of ML systems to identify emerging threats as they develop, without having to be defined prior to deployment. The study utilized three different algorithms; RF, SVM and NN, to identify anomalies in the way a system behaves by tracking file system changes and encryption activities, as well as unusual activity in network traffic. A total of 2850 samples were used for the research, which were sourced from various public repositories of cyber security information, as well as local system logs and simulated network traffic, all of which were normalized, segmented and had features extracted prior to being analyzed. PCA and correlation analysis was also conducted to reduce the number of redundant variables and increase the overall generalizability of the models. All of the algorithms were trained and tested via k-fold cross-validation, after which each model was evaluated based upon accuracy, precision, recall and F1 score. The NN produced the most accurate results (92.4%); however, both the RF (91.2%) and SVM (88.7%) maintained very similar levels of performance, with all of the models achieving high F1 scores. These results supported the use of ML for identifying zero-day ransomware attacks, since the ML models were able to identify these attacks based upon learned patterns of abnormal behavior, as opposed to the reliance on established signatures.

Cen et al. [48] developed Zero-Ran Sniff (ZRS). This is a novel early ransomware detection framework which applies the methodology of zero-shot learning (ZSL) to recognize unknown ransomware attacks prior to the point of encryption. The researchers attempted to address some of the shortcomings associated with current signature-based approaches, as well as other dynamic feature-dependent methods which do not support new/unknown ransomware variants. The dataset was obtained from Sgandurra et al. (2016). The dataset contained 582 examples of ransomware along with 942 examples of benign samples and spanned 11 different ransomware families; including: CryptoLocker, Locker, and TeslaCrypt. The static Portable Executable (PE) header attributes were extracted using the Python library pefile in order to allow for the efficient processing of data. There are two stages of the proposed ZRS method: AE-CAL (Autoencoder-based Core

Attribute Learning) is the first phase, where AE-CAL extracts deep semantic representations from previously known classes; SA-CNN-IS (Self-Attentive Convolutional Neural Network Inference Stage) is the second phase where it classifies the representation based on previously seen classes. The model was trained by utilizing Adam Optimizer and activation functions of type ReLU. The performance of the model was evaluated using metrics of Accuracy, Precision, Recall, and F1-Score. Results of experiments showed a high level of accuracy (96.02%), a high degree of precision (91.49%), a very high level of recall (98.47%) and a high level of F1-Score (96.31%) compared to state-of-the-art methods of Random Forest, KNN and AdaBoost, as well as to deep learning baselines such as VGG16 and ResNet. Also the researchers improved recall of their zero-day model by 2.9%, and reduced feature extraction time to 0.3 seconds.

Fevid et al. [49] proposed a new zero-day ransomware detection approach based on analysis of assembly language bytecode and random forest classification of unknown ransomware family types based solely on their bytecode characteristics. The researchers sought to improve the early-stage detection rate of ransomware by analyzing the opcode level behaviors common to all ransomware families. The researchers assembled an extensive dataset from VirusShare comprising over 1,700 samples including the major ransomware families (LockBit, REvil, BlackMatter and DarkSide) along with benign executables for comparative control testing. The researchers used IDA Pro and Radare2 to decompile each executable and generate opcode frequencies, instruction sequences, entropy values and control flow graphs, which were subsequently converted into vectorized feature formats via n-gram tokenization and principal component analysis (PCA). The Random Forest classifier was comprised of 500 decision trees and its performance was assessed against several metrics: Precision, Recall, F1 score and Accuracy. The developed model demonstrated an accuracy of 93.5%, precision of 0.95, recall of 0.94 and F1 score of 0.945 and demonstrated strong generalizability to previously unseen ransomware samples based on a high ROC-AUC curve. The results also indicated that opcode frequency and instruction sequence pattern were significantly more indicative of ransomware behavior than deep learning techniques such as CNNs while providing better performance at a lower computational cost. However, the authors acknowledged several limitations of the developed model, including reliance upon static features, inability to handle polymorphic ransomware and potential false negative results due to high levels of code obfuscation. They suggested utilizing a combination of static and dynamic analysis techniques, representing control flow as graphs and employing hybrid machine learning architectures to develop a future generation of ransomware detection frameworks with greater adaptability, scalability and real-time responsiveness.

A. Summary of the Discussion of Related Work

In summary, Ransomware detection systems can only be built on datasets to be reliable. In order to train and evaluate the ransomware detection models with the help of ML, several publicly available datasets have been extensively used, including CICandMAL 2017, ISOT, VirusShare randoms dataset, UGRansome, UGRansome1819, and UGRansome-2024.

Table I presents an overall comparison of some of the most common ransomware detection datasets. Each dataset is analyzed based on its features, data distribution, and recognized shortcomings. Most of the existing ransomware datasets have the same issue: synthetic data, suffer from a serious class-imbalance problem or lack newer ransomware.

TABLE I. SUMMARY OF THE RANSOMWARE DETECTION DATASETS AND THEIR KEY FEATURES AND LIMITATIONS.

Author(s)	Dataset(s)	Features	Limitation(s)
[39]	UGRansome1819	Logistics and Network flow	Needs more real-world validation
[26]	UGRansome1819	14 network attributes	Potential overfitting
[27]	UGRansome2024	12 selected features	Imbalanced classes
[32]	UGRansome1819	14 attributes including financial features	Imbalance and lack of diverse ransomware
[28]	UGRansome	Multiple network flow	Balance issues
[25]	UGRansome	14 features including protocol and flags	Class imbalance
[34]	UGRansome	14 features network and transaction	Imbalanced ransomware family
[30]	UGRansome1819	Diverse, protocol, payload, time, etc.	Legacy datasets used as benchmarks are outdated
[31]	UGRansome	Multi-type network and behavior features	High dimensionality
[33]	UGRansome	Network and financial features	limitations in diversity and imbalance
[41]	Android Ransomware Dataset	19 selected network traffic features	lacks real-world variety
[42]	ISOT Ransomware Dataset	Network flow and time-based statistics reduced by PCA	Binary labeling only benign/malicious
[43]	CICAndMal2017	9 key timing and packet-size traffic features	Limited representation of new ransomware
[44]	Custom hybrid network dataset	Extracted CNN features s	Ptential false positives on encrypted flows
[45]	Custom dataset	Flow-level features: packet size, duration,etc..	High computational cost for dual models
[46]	VirusTotal Intelligence	IP, ports, protocol type, etck	scalability for large-scale networks
[47]	Public malware repositories	File system changes, and log features	High CPU/memory cost
[48]	Viursshare dataset	PE header static features and Self-Attentive CNN layers	Windows PE structure
[49]	VirusShare (12 ransomware families)	Opcode frequency, and entropy-based attributes	Dynamic behavior capture
[40]	UGRansome1819	Network flow	Outdated for most recent ransomware trends

IV. METHODOLOGY

A. Research Framework

In this study, we propose a supervised machine learning framework for ransomware detection in network traffic. The overall workflow, shown in Fig. 4, follows these stages which are: data collection, preprocessing, feature engineering, classification, evaluation, and validation.



Fig. 4. Overall methodology framework

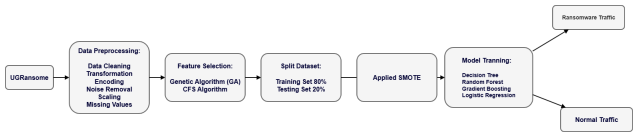


Fig. 5. Overall methodology framework illustrating the six main stages of the proposed ransomware detection pipeline.

B. Experiments

We ran all of our experiments on a machine that had the exact same operating system Ubuntu 22.04, as well as programming language Python 3.10, and also included the hardware Intel Core i9 13th Gen and 64GB DDR5 RAM.

C. Dataset Description and Preprocessing

As shown in Table II, the UGRansome dataset comprises 149,043 network traffic entries of which fourteen (14) numerical and categorical variables have been defined.

TABLE II. SUMMARY OF THE UGRANSOME DATASET

Attribute	Description
Dataset Name	UGRansome (UGRansome1819)
Source	University of Pretoria; combines UGR16 traffic data with ransomware-specific network traces [29]
Total Instances	149,043 network traffic records
Number of Features	14 attributes (13 predictive + 1 target label)
Feature Types	7 categorical (Protocol, Flag, Family, SourceAddress, ExpAddress, IPAddress, ThreatType) and 6 numerical (Time, BTC, USD, Netflow, Port, Clusters)
Class Distribution	Normal: 108,941 (73.09%) | Ransomware: 40,102 (26.91%)
Imbalance Ratio	2.72:1 (Normal : Ransomware)
Missing Values	None detected
Preprocessing	Label encoding for categorical variables; StandardScaler normalization for continuous features; duplicate identifier removal
Class Balancing	SMOTE applied on training set only; 55,072 synthetic ransomware samples generated (0.56 s)
Train/Test Split	80/20 stratified split (119,234 training / 29,809 testing); random seed = 42

This dataset encompasses an array of different network behavior that is representative of both typical activity, as well as ransomware-based activity.

The "Prediction" label has been translated to a binary target value: Normal Network Activity Class 0 and Ransomware Class 1.

About 73% of all instances are classified as normal and approximately 27% as ransomware; these percentages represent a moderate degree of imbalance, which contributed to the decision to apply SMOTE.

Data preprocessing included cleaning the data, converting the categorical variables (Protocol, Flag, Family, SourceAddress, ExpAddress, IPAddress, ThreatType) to encoded format, and normalizing the continuous features using StandardScaler.

No missing values were discovered during the data preprocessing, and duplicate identifier fields were removed, leaving thirteen (13) predictive feature(s).

As shown in the Fig. 5, the implementation steps demonstrate a more detailed view of the data processing and classification.

D. Feature Engineering and Class Balancing

UGRansome database contained 14 attributes covering different characteristics of network traffic, including temporal behavior, type of protocol, financial transactions and threat intelligence. The given attribute that was dropped was the one called Prediction, since it is the target label, and it would cause data leakage to be retained as an input. Following this, the dataset had 13 predictive features.

Among these 13 features, seven of them were categorical (Protocol, Flag, Family, Sender Address, Receiver Address, IP Address and Threat Type). These were coded with labels in order to make them compatible with machine-learning algorithms without expanding the feature space. The other six characteristics (Time, BTC, USD, Netflow\Bytes, Port, and Clusters) were numeric.

The best overall performance was obtained with the full 13-features set at an accuracy of 99.4%, Precision 98.00%, recall 99.9% and a AUC-ROC 99.95%, of the Decision Tree classifier. These findings prove that the features chosen are very informative, non-redundant, and can be used as the most appropriate features to identify ransomware.

Table III presents the statistical validation results for all 13 predictive features.

TABLE III. STATISTICAL VALIDATION OF THE 13 PREDICTIVE FEATURES IN THE UGRANSOME DATASET.

Feature	Correlation	Mutual Information	F-Statistic	p-Value
USD	0.1473	0.4007	9318.68	< 0.001
BTC	0.2426	0.3376	441.38	< 0.001
Port	0.1152	0.0861	27652.33	< 0.001
Flag_encoded	0.0786	0.3016	2005.54	< 0.001
Clusters	0.3005	0.0513	2613.02	< 0.001
Netflow_Bytes	0.1313	0.2497	1280.96	< 0.001
ExpAddress_encoded	0.0474	0.1008	1479.71	< 0.001
SourceAddress_encoded	0.0543	0.0474	936.89	< 0.001
Threats_encoded	0.0077	0.0177	886.57	< 0.001
Time	0.3956	0.0153	17.19	< 0.001
IPaddress_encoded	0.0107	0.0160	827.12	< 0.001
Protocol_encoded	0.0923	0.0140	8.01	0.005
Family_encoded	0.0769	0.0209	336.14	< 0.001

To prevent any kind of data leakage, the data was first divided into training and test sets (20% data as test and 80% data as train) and then SMOTE was applied on the training set only after the split. Likewise, the StandardScaler was only fitted to the training fold, then applied to the training and test folds.

Before oversampling, there are 119,234 records in the training partition (87,153 normal and 32,081 ransomware samples). The training set was completely balanced with 87,153 samples of each class (174,306 total) and synthetic ransomware samples

were generated in 0.56 s with the application of SMOTE. A test set of 29,809 samples was held out and only used for final evaluation.

E. Model Training and Validation

The dataset was divided into two parts by employing an 80/20 stratified split: 119,234 samples for training and 29,809 for validation. All feature scaling occurred post-split (0.03 s).

A fixed random seed (42) was used for all stochastic operations to ensure reproducibility. Four supervised classifiers were trained using `scikit-learn` (v1.3): Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB). The hyperparameters used for each model are listed in Table IV; default scikit-learn values were retained where not specified.

TABLE IV. HYPERPARAMETER CONFIGURATION

Model	Hyperparameters
LR	solver = `lbfgs`, C = 1.0, max_iter = 1000
DT	criterion = `gini`, max_depth = None
RF	n_estimators = 100, max_depth = None
GB	n_estimators = 100, learning_rate = 0.1, max_depth = 3
SMOTE	k_neighbors = 5, sampling_strategy = `auto`

Training times on the balanced data were: DT = 0.37 s, RF = 0.63 s, GB = 13.75 s, LR = 2.07 s. All trained models and preprocessing pipelines were saved for downstream evaluation.

V. EVALUATION METRICS

To evaluate the performance of the developed ransomware detection models, we will use five common evaluation criteria to measure the accuracy of the models' ability to recognize ransomware as well as to differentiate it from normal internet traffic. These five are:

1) Accuracy: It reflects the quality of its predictions, and the model classifies accurately by following the formula given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2) Precision: This measures how many actual instances of ransomware were detected by the model by following the formula given below:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3) Recall: It is a measure of how well the model identifies true cases of ransomware from all the possible ransomware instances that exist by following the formula given below:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4) F1-Score: This metric gives you a single number to assess both the number of False Positives (FP) and False Negatives (FN) of the model by following the formula given below:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5) AUC-ROC: This value measures how good the model is at separating normal and ransomware distributions into normal and ransomware traffic distributions:

$$\text{AUC-ROC} = \int_0^1 TPR(FPR)\, d(\text{FPR}) \tag{5}$$

Here, TP, TN, FP, and FN refer to the counts of *True Positives*, *True Negatives*, *False Positives*, and *False Negatives*, respectively.

VI. RESULTS AND ANALYSIS

The implementation used `scikit-learn` and `imbalanced-learn` to train models; to perform cross-validation; and to create synthetic samples through SMOTE sampling.

A. Overview of Experimental Evaluation

UGRansome dataset was used to test four trained machine learning algorithms, namely, Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) to identify ransomware in network traffic.

Two experimental were conducted: first one, imbalanced data and the second, balanced data (with the help of SMOTE). This two-way comparison allowed to fairly compare the effect of class imbalance on detection accuracy, and the use of synthetic oversampling to enhance the detection of minority (ransomware) samples without affecting overall accuracy.

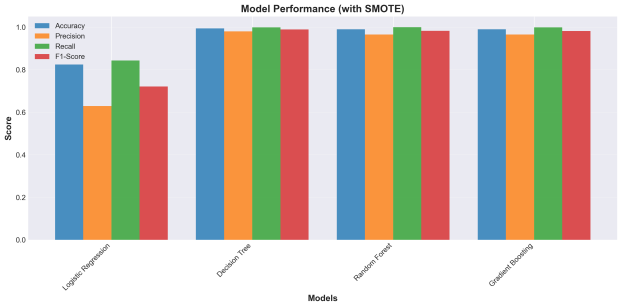


Fig. 6. Performance comparison of all models on the SMOTE-balanced dataset.

As shown in Fig. 6, the Decision Tree had the highest detection accuracy (99.4% and recall (99.9%). On the other hand, Logistic Regression did not perform as well 82.4%, which shows that it is not effective in capturing the complex, non-linear feature interactions that are likely to be in captured encrypted or polymorphic network traffic.

B. Model Performance Comparison

Table V summarizes that the Decision Tree displayed the best overall performance having 99.4% accuracy, 98.0% precision, 99.9% recall, F1-score of 98.9%, and an AUC-ROC of 99.95%. The RF and GB models also had a high level of performance (accuracy of more than 99%), indicating the potential strength of ensemble models.

In contrast, the Logistic Regression obtained an accuracy of 82.4% with an AUC of 89.9%, thus being incompatible with ransomware detection.

TABLE V. PERFORMANCE COMPARISON OF SUPERVISED MODELS (SMOTE-BALANCED DATASET).

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
DT	99.40	98.00	99.90	98.90	99.95
RF	99.00	96.50	100.00	98.20	99.98
Gradient Boosting	99.00	96.60	99.90	98.20	99.97
Logistic Regression	82.40	62.90	84.30	72.10	89.90

C. Confusion Matrix Analysis

Fig. 7 and Fig. 8 demonstrate the model behavior balancing with and without SMOTE. This model before balancing, the model have good results precision, but can't classify many of the ransomware examples, 43 false negatives. When applying SMOTE, this model (DT) had only eight false negatives resulting in an 81% improvement in ransomware detection and very little loss in precision. This enhancement shows the effectiveness of teaching the minority class, in order to teach the model itself, more efficiently rarely attack patterns.

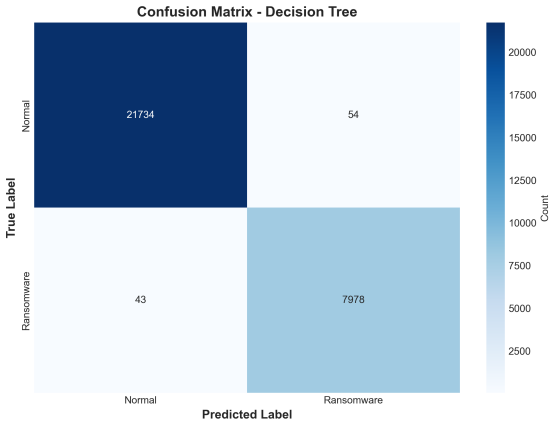


Fig. 7. Confusion matrix of the Decision Tree on the original (unbalanced) dataset.

D. ROC Curve Analysis

The ROC curve illustrated in Fig. 9 and Fig. 10 are almost perfectly classified with AUC of 99.95% or higher with either of the data points. The SMOTE-balanced showed some consistent improvement in the detection performance at lower thresholds with the consequence that fewer false negatives are obtained at no cost to specificity. Conversely, the Logistic Regression model gave a much flatter curve (AUC = 89.9%) that gave a relatively low discriminatory capability between

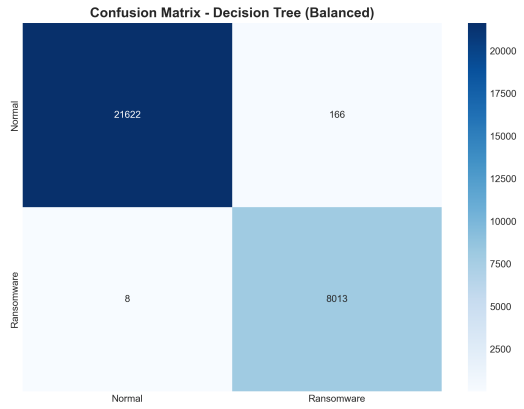


Fig. 8. Confusion matrix of the Decision Tree after SMOTE balancing.

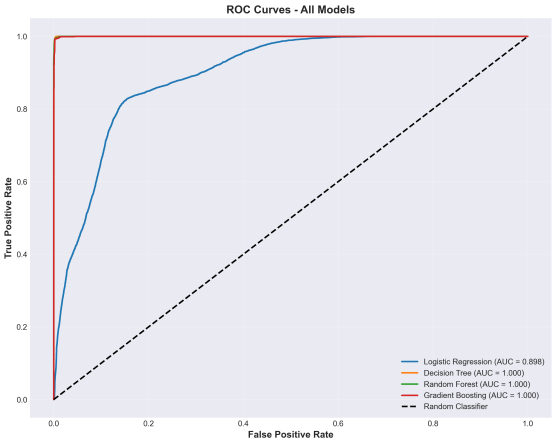


Fig. 9. ROC curves for all models on the unbalanced dataset.

normal and ransomware traffic. To further explore the contribution of various features, two feature selection techniques were compared with the baseline (all 13 attributes): Correlation Based Feature Selection (CFS) and Genetic Algorithm (GA).

E. Comparison of Feature Selection Methods on the Decision Tree Model

As shown in Table VI, it illustrates that the complete feature set was the most accurate and least time-consuming. In fact, CFS substantially decreased dimensionality, it caused a decrease in accuracy by 4.23% and falsely negative results. GA was also able to compete at eight features (99.07% all accuracy) with approximately one second of computation overhead. On the whole, as the dataset already includes a rather small amount of features (13), it made it effective and efficient to take all of them as the ransomware features.

TABLE VI. COMPARISON OF FEATURE SELECTION METHODS ON THE DECISION TREE MODEL.

Method	Features	Acc.	Prec.	Rec.	F1	Total(s)
All Features	13	**99.67**	**99.33**	**99.46**	**99.40**	**0.13**
GA	8	99.07	97.20	99.41	98.29	1.03
CFS	3	95.44	90.43	92.87	91.63	0.37

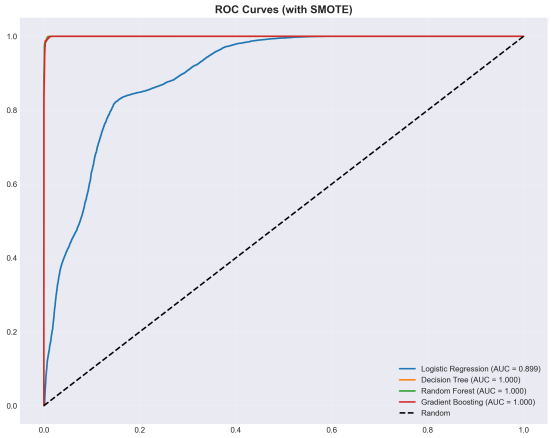


Fig. 10. ROC curves for all models after SMOTE balancing.

F. Comparative Analysis: Original vs. SMOTE Pipeline

Fig. 11 shows comparative analysis. The use of SMOTE decreased precision (-1.36%), accuracy (-0.26%) and recall increased by +0.44This trade-off is positive in the context of cybersecurity, since dedicating more time to identifying attacks is much more important than preventing some additional false signals. This proves that training is strengthened and more trusted in the case of SMOTE-enhanced ransomware detection.

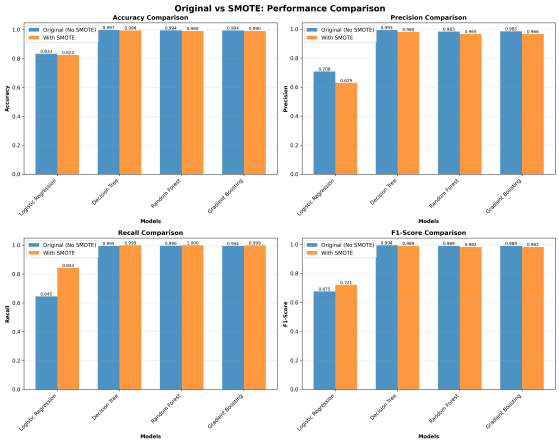


Fig. 11. Comparison of model metrics between original and SMOTE-balanced pipelines.

G. Computational Efficiency

As shown in Fig. 12 and Table VII, Decision Tree train 0.37s to 4ms with a throughput rate of approximately 7.5 million samples per second. Logistic Regression achieved greater throughput (13.7M/s) its accuracy is less than ideal. Random Forest and Gradient Boosting moderate latency (63ms,34ms) and can be fit in real-time security appliances.

H. Summary of Results and Analysis

In short, the Decision Tree model, which was trained on the SMOTE-balanced dataset, has the highest overall balance between precision, recall and processing speed. It presented

TABLE VII. Training and Inference Efficiency of Evaluated Models.

Model	Training Time (s)	Test Time (ms)	Throughput (samples/s)
Decision Tree	0.37	4	7.5M
Random Forest	0.63	63	471K
Gradient Boosting	13.75	34	874K
Logistic Regression	2.07	2	13.7M

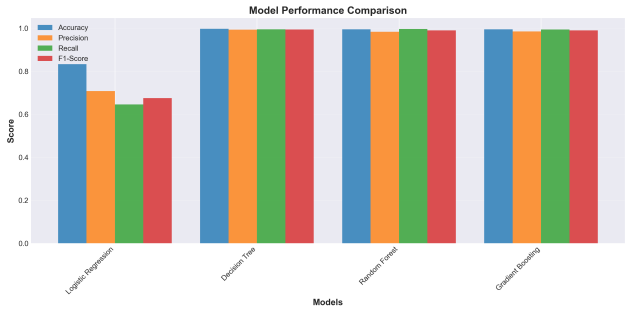


Fig. 12. Comparative computational performance (training and testing times).

almost an ideal AUC of 99.95%, and it is also highly interpretable and efficient in real-time detection. These findings confirm that transparency and rule-based models can significantly enhance the performance and high-quality outcomes of IDS in relation to ransomware attack-resistance.

In addition, All Features also supported this conclusion as the highest accuracy (99.65%) was obtained using all 13 attributes with the least amount of runtime. This validates the move to maintain the entire feature set within the ultimate detection model.

VII. Discussion

The results of the experiments clearly illustrate that supervised learning models, particularly the Decision Tree model applied to the SMOTE balanced traffic data are very effective in the detection of malware associated with ransomware applications. As we discussed in the pervious sections, the model Decision Tree shows impressive performance metrics of 99.40% accuracy, 98.0% precision and 99.9% recall confirming the model's strong ability to identify complex behaviors associated with ransomware. With an equally impressive AUC R - ROC metric of 99.95%, the Decision Tree model can easily distinguish between benign normal traffic and malicious traffic. Although there are ensemble machine learning algorithms like Random Forests and Gradient Boosting which have similar accuracy, the Decision Tree algorithm has the unique advantage of speed of computation, thus making it suitable for cyber defense applications.

The analysis of feature importance reflected certain key features such as the type of protocol used, the presence of certain flags, and source/destination addresses as well as transaction-related features such as the USD value and clustering information of the traffic data. These were thus the important features in malware detection associated with ransomware behavior. This result is in accordance to much earlier experimental work that has linked unusual protocol-type activity or behavior shown by network traffic and the

suspicious transactions associated with ransomware and the theft of data. The application of SMOTE for effectively balanced datasets shows promise for the considerable decrease of false negatives of around 81% and the detection of even rare types of ransomware without loss of accuracy. This is an important tradeoff in intrusion detection applications where the loss of an attack results in much greater consequences than the triggering of a few additional alerts. In the comparison of feature selection methods performed, it was found that all 13 features produced the best results by achieving 99.67% accuracy with an insignificant increase in runtime. Therefore, the decision was made to keep all features within the model, as an appropriate, well defined and complete feature collection aids in sustaining accuracy without additional expense.

From a pragmatic view point, the Decision Tree model was found to produce the best combination of accuracy, transparence and run-time. It was trained very quickly and thus completed training in 0.37 seconds and produced predictions of approximately 4 ms for each prediction, with the ability to process more than 7.5 million samples a second. Such efficiency would easily allow for integration of the model within practical situations such as Security Information and Event Management (SIEM) OR NIDS where decisions need to be made rapidly and transparency is the requirement. It can be seen that our study has found that lightweight, interpretable models lend themselves to improvement the data they are trying to process and, through intelligent use of feature engineering can rival the larger and more complex algorithms and be much more serviceable to real time processing.

There are numerous possible avenues that our work could take in the future. Enlarging our data to include more families of ransomware and real network traffic data would help the model be more discriminating and able to detect new attacks. Future work could assess the logic of using interpretable models in conjunction with other ensemble learning methods or deep learning methods, to form resilience to zero day attacks or adversarial attacks. Other avenues of study could include features that use explainability tools such as SHAP or LIME which would help Security Analysts determine the reasoning behind the model decisions taken.

Finally, deployment of such an integrated detection mechanism could be ideal in applications where the distributed nature of processing or streaming nature of the data, as would be the case using Apache Flink or Google Cloud Dataflow. This enables testing of the algorithms in real network conditions of high speed and large scale that would be the requirement in real industrial circumstances. Integrating within the detection algorithms other data sources, for example, telemetry data from endpoints or block chain based threat indicators, could assist greatly in providing detection within a Zero Trust Security Framework. It is hoped that our work will continue along these routes and move this work along from a successful experimental prototype to a practical, adaptive and transparent model of intrusion detection that is capable of combating the ever evolving and changing face of the ransomware threat we face.

We deliberately scoped this study to classical supervised classifiers in order to isolate the effect of SMOTE balancing and feature selection under identical, fully reproducible conditions. Recent deep learning approaches reported on the same UGRansome dataset including the BERT-based model of Rahman et al. [26] (99.21%), the TEVEInet transformer VAE hybrid of Fan et al. [31] ($F_1 = 99.26\%$), and the SAE–LSTM architecture of Nkongolo et al [35] (98.5%) report accuracy that is comparable to, or marginally lower than, our Decision Tree result (99.40%), but at substantially higher computational cost and reduced interpretability. Our framework reaches equivalent accuracy with an inference latency of 4 ms per batch and a throughput of 7.5 M samples/s, suggesting that for the UGRansome dataset the marginal accuracy gain offered by neural architectures does not justify their deployment cost in real-time IDS pipelines. A direct empirical comparison with CNN–LSTM and transformer baselines under our exact preprocessing pipeline is identified as future work in the Limitations section.

A. Limitations

While the claimed success is high, there are limitations to this study that should be recognized, as they guide the interpretation of the current results, and future work in this field.

1) Single dataset dependency: We evaluate only using the UGRansome dataset. Although UGRansome is a prominent and publicly available malware traffic dataset that features realistic event trace of notorious ransomware its traffic characteristics may not be typical for real enterprise environments [29]. Variations in network topology, traffic volume, protocol distribution and user behavior for different organizations may impact model generalization. Validation on other datasets such as CICAndMal2017 or on live traffic captures would enhance the external validity of the proposed framework.

2) SMOTE-Induced overfitting risk: The use of SMOTE to tackle the class imbalance issue creates synthetic samples by interpolating between the existing minority class samples. This can lead to samples that are extremely similar to the training data, which can lead to an overestimation of performance due to overfitting to synthetic patterns rather than learning generalizable ransomware signatures. Although the use of SMOTE was restricted to the training partition (it was never used in the test set), and the high AUC-ROC scores across all the tree-based models indicate high discrimination power, further work is required in order to consider alternative balancing techniques (such as ADASYN, borderline-SMOTE, or cost-sensitive learning) in order to reduce this risk.

3) Binary classification scope: The existing framework conducts binary classification (normal vs. ransomware) and fails to differentiate ransomware families and detect multi-stage attack progression. In practice, the ability to group specific ransomware families (e.g. WannaCry, Locky, CryptoLocker) would make it possible to take more targeted incident response actions. Future extensions should explore multi-class classification with labeling at the family level, and multi-stage detection of the ransomware lifecycle, from initial infection to encryption to command-and-control communication.

4) Encrypted traffic challenges: Modern ransomware often uses the SSL (HTTPS, TLS, etc.) for hiding the command and control communications as well as for concealing data exfiltration activities. The UGRansome dataset is not specific to encrypted traffic scenarios and the features used in this

study (protocol flags, transaction values, network flow characteristics) may have had reduced discriminative power when payloads are encrypted. Addressing this limitation, encrypted traffic analysis techniques like TLS fingerprinting, JA3/JA3S hashing or metadata based analysis working on connection level attributes instead of payload content need to be incorporated.

5) Absence of deep learning and hybrid approaches: This study is only limited to classical supervised learning algorithms. While the results show that tree-based models show excellent results on the UGRansome dataset, recent developments in deep learning (e.g. CNN-RF hybrids, Transformer-based models, and LSTM-autoencoder architectures have shown promising results for the detection of more sophisticated and evasive ransomware variants. Including such techniques in future comparative research may help understand if further accuracy improvements are possible, especially with polymorphic or fileless ransomware edge cases [26], [44].

6) Limited real-time deployment validation: While the computational efficiency metrics suggest the Decision Tree model can reach throughputs in excess of 7.5 million samples per second, these results were derived in an offline environment. In a real-world network, other factors come into play such as the time taken to capture packets, extract features, the memory footprint of the application and other simultaneous processing tasks. A comprehensive assessment of system throughput, latency, and CPU/memory usage under realistic traffic loads remains necessary to validate practical deployability.

7) Default hyperparameters: All classifiers were trained using default scikit-learn hyperparameters with no optimization using grid search or random search, or Bayesian methods. While the good baseline results indicate that the feature representations are very discriminative, further improvements may be achieved through systematic hyperparameter tuning, especially for Gradient Boosting (where learning rate, tree depth and number of estimators interact non-linearly) and Logistic Regression (where the strength of regularization affects the decision boundary).

8) Dataset bias and result interpretation: The very high accuracy scores (exceeding 99% for the tree-based models) should be interpreted with care. Such performance may be partly a reflection of aspects of the UGRansome dataset itself, such as there being clear separation between class distributions or that there is availability of strongly predictive features (perhaps ones that may not be available in all real-world network environments).

VIII. Conclusion and Future Work

This study proposed a supervised ML technique to detect ransomware in network traffic using the UGRansome dataset. The complete workflow from data preprocessing and normalization to model training, balancing and evaluation was carefully undertaken to ensure its reproducibility and practicability for real-world cyber-security applications.

The proposed framework combines good preprocessing and balancing methodologies that help in ensuring the reliability and consistency of model performance. In this study we applied the four supervised algorithms: DT, (RF), (GB) and (LR) on the original imbalanced dataset and on its SMOTE balanced version. There were clearly observable differences between the balanced and unbalanced versions with respect to recall (improvement of approximately 81%) and also the benefits which followed from a large reduction in false negatives at the point of detecting ransomware attacks.

Of the models tested, the decision tree was the best overall performer with a good compromise on each of accuracy,and speed. Achieving an accuracy of 99.4% with a recall of 99.90% AUC-ROC area under the receiver operating characteristics curve of 99.95%, indicated a very positive capability of separation of normal and ransomware traffic. The model performed its training in less than 1 second, made for predictions largely in negligible time and hence is suitable for real time use in security systems. These results seem to indicate that simple and explainable models can perform to good effect and quality as the more complicated alternatives, whilst remaining of low weight suitable for speedy execution.

A comparison between Genetic Algorithm (GA) and Correlation-Based Feature Selection (CFS) showed that using all 13 features yielded the best accuracy (99.67%), precision (99.33%), recall (99.46%), and F1-score (99.40%) with the lowest computational overhead compared to the other techniques.

This result would indicate that in situations, where datasets are smaller, cleaner and compact in their construction that by employing provision of the whole range of features a superior result can be attained without supplementary feature reduction being necessary.

It can be seen that this study offers a framework for a system and approach to ransomware detection in real network environments which is highly efficient and user friendly, accurate and interpretable. With high performance, low latencies and good transparency this has strong claims to be introduced into IDS (Intrusion Detection System) and SIEMs (Security Intelligence Enabling Modules).

Future work will be directed into extending the dataset to include more ransomware families, hybrid and explainable AI models together with deployment of the technology in either streaming or distributed environments to increase scalability, flexibility and confidence of automated ransomware defense.

Funding

This work was funded by the King Faisal University, Saudi Arabia [Project No. GRANT KFU262639].

Acknowledgment

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. GRANT KFU262639].

Conflicts of Interest

All authors declare no conflict of interest.

Author's Contributions

All authors equally contributed.

REFERENCES

[1] Y.-L. Wan, J.-C. Chang, R.-J. Chen, and S.-J. Wang, "Feature-selection-based ransomware detection with machine learning of data analysis," in *2018 International Conference on Computer Communication and Multimedia (CCOMS)*, Apr. 2018.

[2] K. Schmaltz, S. C. Thompson, D. F. da C. M. Mendes, and J. Carvalho, "Robust defense mechanisms against adversarial ransomware attacks: Implementing a universal network-level detection filter," Sep. 2024.

[3] S. Ahmad, Z. Zulkifli, N. H. Nasarudin, M. Imran, and M. Ariff, "A recent systematic review of ransomware attack detection in machine learning techniques," in *2023 International Conference on Artificial Intelligence and Data Science (AIDAS)*, Sep. 2023.

[4] A. Abcam, L. Cranshaw, C. Worthington, and J. Vandenberghe, "Behavioral pattern analysis for real-time detection of ransomware attacks," Jan. 2025.

[5] P. Novak, P. Kaura, V. Oujezsky, and T. Horvath, "Ransomware file detection using hashes and machine learning," in *2023 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Jan. 2023.

[6] K. Higuchi and R. Kobayashi, "Rofbsa: Real-time backup system decoupled from ml-based ransomware detection," arXiv preprint, Apr. 2025.

[7] L. Chen, C.-Y. Yang, A. Paul, and R. Sahita, "Towards resilient machine learning for ransomware detection," *Proceedings of KDD Workshop on Security and Privacy*, pp. 1–10, 2019.

[8] C. Khoza, G. M. Komba, and S. P. Maswikaneng, "An enhanced ransomware defense strategy through behavior-based detection using machine learning algorithms," in *2024 IEEE 8th Conference on Energy Internet and Energy System Integration*, Jan. 2024.

[9] P. Veerasingam, S. A. Razak, A. F. A. Abidin, M. A. Mohamed, and S. D. M. Satar, "Intrusion detection and prevention in smes using suricata," *Malaysian Journal of Computing and Applied Mathematics*, vol. 6, no. 1, pp. 21–30, 2023.

[10] N. Rani and S. V. Dhavale, "Leveraging machine learning for ransomware detection," arXiv preprint, Jun. 2022.

[11] V. Kumar and O. P. Sangwan, "Signature based intrusion detection system using snort," *International Journal of Computer Applications in Information Technology*, vol. 1, no. 3, pp. 35–41, 2012.

[12] S. Chakrabarti, M. Chakraborty, and I. Mukhopadhyay, "Study of snort-based ids," in *Proceedings of International Conference & Workshop on Emerging Trends in Technology*, 2010, pp. 43–47.

[13] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, "An intelligent intrusion detection system (ids) for anomaly and misuse detection in computer networks," *Expert Systems with Applications*, vol. 29, no. 4, pp. 713–722, 2005.

[14] M. Al-Zewairi *et al.*, "Multi-stage enhanced zero trust intrusion detection system for unknown attack detection in internet of things and traditional networks," *ACM Transactions on Privacy and Security*, 2025.

[15] A. O. Almashhadani, M. Kaiiali, S. Sezer, and P. Okane, "A multi-classifier network-based crypto ransomware detection system: A case study of locky ransomware," *IEEE Access*, vol. 7, pp. 47 053–47 067, 2019.

[16] S. Sivalmi, Jayashree, and H. Noor, "The dual role of ai in cyber security: Enhancing ransomware and defending against attacks," *Journal of Computer and Communication Systems*, vol. 1, no. 3, pp. 20–30, 2025, discussion on ransomware evolution, impact including economic/social, and attack strategies, alongside AI-powered intrusion detection systems and attack countermeasures.

[17] M. Benmalek, "Ransomware on cyber-physical systems: Taxonomies, case studies, security gaps, and open challenges," *Internet of Things*, 2024, comprehensive overview of ransomware threats, taxonomies, and implications on cyber-physical systems, highlighting attack vectors, organizational impacts, and mitigation challenges.

[18] W. Z. A. Zakaria, M. F. Abdollah, O. Mohd, S. M. W. Mohamed, M. Yassin, and A. Ariffin, "Rentaka: A novel machine learning framework for crypto-ransomware pre-encryption detection," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, pp. 378–387, 2022, details ransomware definitions, lifecycle, and ML-based early detection with emphasis on cryptographic ransomware behavior and attack vectors.

[19] M. Benmalek, "Ransomware on cyber-physical systems: Taxonomies, case studies, security gaps, and open challenges," *Internet of Things and Cyber-Physical Systems*, Jan. 2024.

[20] Syed Shameem, "Estimating malware impact on network traffic analysis by using wireshark," vol. 20, no. 7, pp. 965–972. [Online]. Available: https://journal.esrgroups.org/jes/article/view/3477

[21] B. Dodiya and U. K. Singh, "Malicious traffic analysis using wireshark by collection of indicators of compromise," vol. 183, no. 53, pp. 1–6. [Online]. Available: http://www.ijcaonline.org/archives/volume183/number53/dodiya-2022-ijca-921876.pdf

[22] G. Pimenta Rodrigues, R. De Oliveira Albuquerque, F. Gomes De Deus, R. De Sousa Jr., G. De Oliveira Júnior, L. García Villalba, and T.-H. Kim, "Cybersecurity and network forensics: Analysis of malicious traffic towards a honeynet with deep packet inspection," vol. 7, no. 10, p. 1082. [Online]. Available: https://www.mdpi.com/2076-3417/7/10/1082

[23] I. Sembiring, S. , A. Iriani, J. V. B. Ginting, and J. A. Ginting, "A novel approach to network forensic analysis: Combining packet capture data and social network analysis," vol. 14, no. 3. [Online]. Available: http://thesai.org

[24] M. Williams, R. Morales, K. Johnson, G. Martinez, and J. Bennett, "Entropy-based network traffic analysis for efficient ransomware detection." [Online]. Available: https://doi.org/10.36227/techrxiv.172840776.66718131/v1

[25] I. Chaudhary and S. Adhikari, "Ransomware detection using machine learning: A comparative study," *Researcher CAB*, vol. 10, no. 1, pp. 96–114, 2024.

[26] M. Abdur Rahman *et al.*, "Machine learning models for ransomware detection using ugransom1819," *Computers*, 2025.

[27] P. Azugo, H. Venter, and M. W. Nkongolo, "Ransomware detection and classification using random forest: A case study with the ugransome2024 dataset," *arXiv preprint arXiv:2404.12855*, 2024.

[28] S. J. Nhlapo and M. N. W. Nkongolo, "Zero-day attack and ransomware detection," 2024. [Online]. Available: https://arxiv.org/abs/2408.05244

[29] M. Nkongolo, J. P. van Deventer, and S. M. Kasongo, "Ugransome1819: A novel dataset for anomaly detection and zero-day threats," *Information*, vol. 12, no. 10, p. 405, 2021.

[30] A. Alraizza and A. Algarni, "Ransomware detection using machine learning: A survey," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 143, 2023.

[31] H. Fan, W. Le, Z. Jia, H. Zhao, C. He, H. Jiang, Z. Hu, X. Lv, J. Yuan, and X. Huang, "Teveinet: A deep feature extraction network for anomaly detection," Preprint, SSRN, 2025, https://ssrn.com/abstract=5325605.

[32] S. R. Zahra, "Ransomware detection using ensemble machine learning models with ugransom1819," *Journal of Cybersecurity Research*, 2022.

[33] A. A. Alhashmi, A. A. Darem, A. B. Alshammari, L. A. Darem, H. K. Sheatah, and R. Effghi, "Ransomware early detection techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14 497–14 503, 2024.

[34] P. Yan, T. T. Khoei, R. S. Hyder, and R. S. Hyder, "A dual-stage ensemble approach to detect and classify ransomware attacks," in *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, pp. 781–786. [Online]. Available: https://ieeexplore.ieee.org/document/10754695/

[35] M. N. W. Nkongolo, "Ensemble learning and genetic algorithm for the detection of novel network threat anomaly using the ugransome dataset," Ph.D. thesis, University of Pretoria, Faculty of Engineering, Built Environment and Information Technology, 2023, supervisor: Dr. Jacobus Philippus van Deventer; Co-supervisor: Dr. Sydney Mambwe Kasongo.

[36] M. N. W. N, "Zero-day vulnerability prevention with recursive feature elimination and ensemble learning," Cryptology ePrint Archive, Paper 2023/1843, 2023. [Online]. Available: https://eprint.iacr.org/2023/1843

[37] S. K. D.R.V.A, R. V, K. Rayudu, V. A, T. B, and S. K. R, "Enhancing ransomware detection in cybersecurity: A comprehensive ensemble approach," *Journal of Electrical Systems*, vol. 20-10s, pp. 5222–5232, 2024.

[38] B. Torky, "Ensemble methods for the anomaly detection in enterprise systems," Master's thesis, Rochester Institute of Technology, Department of Graduate Programs Research, RIT Dubai, 2023, capstone, Master of Science in Professional Studies Data Analytics.

[39] A. Mohamed, M. Mohamed, and M. Ata, "Zero-day exploits detection with adaptive wavepca-autoencoder (awpa): Adaptive hybrid exploit detection," 2025, submitted Manuscript.

[40] E. N. Mutombo and M. W. Nkongolo, "Blockchain security for ransomware detection using machine learning and the ugransome dataset," *arXiv preprint arXiv:2407.16862*, 2024. [Online]. Available: https://arxiv.org/abs/2407.16862

[41] A. A. Ahmed, A. Shaahid, F. Alnasser, S. Alfaddagh, S. Binagag, and D. Alqahtani, "Android ransomware detection using supervised machine learning techniques based on traffic analysis," *Sensors*, vol. 24, no. 189, 2024.

[42] A. Srivastava, N. Kumar, A. Handa, and S. K. Shukla, "Ransomware detection based on network behavior using machine learning and hidden markov model with gaussian emission," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2023.

[43] C. Manzano, C. Meneses, and P. Leger, "An empirical comparison of supervised algorithms for ransomware identification on network traffic," in *2020 IEEE ANDESCON*, 2020.

[44] L. Woo, S. Tan, T. Tan, and G. Lao, "A hybrid approach to ransomware detection using convolutional neural networks and random forests on network traffic patterns," in *2024 International Conference on Cyber Security (ICCS)*, 2024.

[45] I. Batalov, P. Smirnov, and A. Lebedev, "Ransomware detection via network traffic analysis using isolation forest and lstm neural networks," *Journal of Cybersecurity and Digital Forensics*, vol. 12, no. 3, pp. 145–160, 2024.

[46] S. Mehrban and E. Geransayeh, "Ransomware detection using machine learning algorithms and network traffic features," *Computer Communications*, vol. 213, pp. 190–204, 2024.

[47] J. Brinkley, S. Alvarez, and W. Zhang, "Machine learning-based intrusion detection for zero-day ransomware threats," *IEEE Access*, vol. 12, pp. 73 210–73 224, 2024.

[48] Y. Cen, R. Zhang, M. Liu, and Y. Hu, "Zero-ran sniff: Zero-shot learning for early detection of zero-day ransomware," *Computers & Security*, vol. 145, pp. 103–119, 2024.

[49] E. Fevid, C. Walsh, and L. Russo, "Zero-day ransomware detection via assembly language bytecode analysis and random forest classification," *TechRxiv Preprint*, September 2024, available at https://doi.org/10.36227/techrxiv.17256576.v1.

Domain Knowledge-Enhanced Welding Robot Path Planning Algorithm

Ming Liu, Peng Shao*
School of Mathematics and Information Science, Chaoyang Normal University, Chaoyang, China

***Abstract*—Welding Robot Path Planning (WRPP) is a core technology for welding automation that seeks an optimal path that satisfies the requirements of the welding process in complex obstacle environments. However, existing algorithms generally lack the integration of knowledge in the welding domain and exhibit insufficient adaptability to complex working conditions. To address these issues, this study proposes a Domain Knowledge-Enhanced Welding Robot Path Planning Algorithm (DKE-WRPP) equipped with a pluggable knowledge fusion framework, which is validated on three representative path planning algorithms: Artificial Bee Colony (ABC), Genetic Algorithm (GA), and Optimal Rapidly-exploring Random Tree (RRT*). Specifically, we first employ Large Language Models (LLMs) to extract domain knowledge such as welding processes and safety distances, and generate standardized knowledge vectors via semantic encoding using a pre-trained language model. Then, a Knowledge Enhancement Module (KEM) is constructed to deeply fuse knowledge features and path geometric features through an attention mechanism, and adaptively update the cost functions of the three baseline algorithms, realizing low-intrusive coupling between domain knowledge and planning algorithms. Finally, experiments in a 300×300 grid environment demonstrate that, compared with traditional algorithms, the knowledge-enhanced algorithms reduce the convergence iterations by more than 33% on average and significantly improve path smoothness. The results fully verify the effectiveness of domain knowledge enhancement and the universality of the pluggable framework, providing an efficient and stable solution for welding robot path planning in complex working conditions.**

***Keywords*—Path planning; intelligent optimization algorithm; welding robot; Large Language Models**

I. Introduction

With the intelligent upgrading of modern manufacturing, robotic technology has evolved from simple industrial automation equipment to sophisticated intelligent autonomous robots. Initially, robots were mainly used for highly repetitive mechanized tasks in industrial production [1]. Driven by advances in computing power, integrated circuits, and sensor technology [2], robots have acquired environmental perception and autonomous decision-making capabilities. Today, robotic applications have expanded from traditional industrial manufacturing to medical, agricultural, welding, and other fields, becoming an indispensable key technology across industries [3].

Welding Robot Path Planning(WRPP) [4], [5], [6], [7] is a core technology of robot intelligence. Essentially, it seeks an obstacle-free optimal path for a welding robot between a start point and a target point under specific optimization criteria. An efficient and reasonable path planning scheme can significantly improve welding efficiency, reduce equipment motion wear, and ensure safe and stable production.

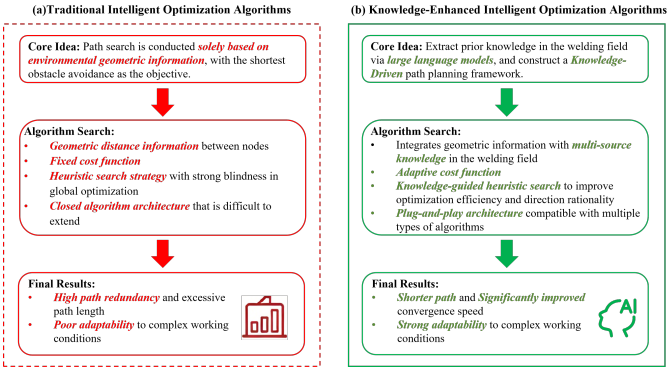


Fig. 1. Comparison of traditional and knowledge-enhanced intelligent optimization algorithms.

In recent years, extensive research [8], [9], [10], [11] has been conducted on welding robot path planning worldwide. Inspired by swarm intelligence and stochastic search, researchers have proposed the Artificial Bee Colony (ABC) algorithm [12], which performs collision-free path optimization through distributed search but suffers from strong search blindness in early iterations, leading to slow convergence. The Genetic Algorithm (GA) [13] was then introduced to achieve global path search via crossover, mutation, and selection mechanisms, yet it still faces issues such as premature convergence and insufficient path smoothness. Subsequently, intelligent optimization methods such as the Optimal Rapidly-exploring Random Tree (RRT*) [14] algorithm emerged. With probabilistic completeness and asymptotic optimality, these methods have partially alleviated the problem of finding optimal paths in complex environments.

Despite progress in geometric path planning, the limitations of these methods have become increasingly prominent. Existing algorithms generally **lack integration with welding domain knowledge** and **have poor adaptability to complex working conditions**, which increases computational overhead and severely degrades model performance [15], [16], [17], [18], [19], [20]. As shown in Fig. 1(a), traditional intelligent optimization algorithms perform path search solely based on environmental geometric information, with a closed architecture relying only on geometric distance between nodes. In multi-constrained practical welding scenarios, such algorithms tend to generate redundant paths and low process matching, failing to meet the demands of high-precision automated welding.

*Corresponding author

Notably, Large Language Models (LLMs) [21] offer a new approach for overcoming the knowledge bottleneck in traditional path planning, owing to their superior capabilities in knowledge extraction and multi-source information fusion. As illustrated in Fig. 1(b), a knowledge-driven path planning framework can be built based on LLMs. By fusing geometric information with multi-source welding domain knowledge, it effectively mitigates insufficient domain knowledge and poor adaptability to complex working conditions. Meanwhile, an adaptive cost function and pluggable architecture ensure compatibility with various baseline algorithms, exhibiting strong generality and helping to improve performance in welding robot path planning.

To solve the above problems, this study proposes a Domain Knowledge-Enhanced Welding Robot Path Planning Algorithm (DKE-WRPP). This algorithm combines LLMs with traditional intelligent optimization algorithms to build a knowledge-driven path planning framework, which effectively enhances the adaptability and optimization efficiency of the algorithm in complex welding conditions. Specifically, we first extract welding process, safety distance, and other domain knowledge using LLMs, then perform semantic encoding via the pre-trained language model BERT to generate standardized knowledge vectors. Next, we construct a Knowledge Enhancement Module (KEM) that deeply fuses knowledge features with path geometric features through an attention mechanism, and adaptively updates the cost functions of three mainstream path planning algorithms—ABC, GA, and RRT*—to achieve low-intrusive, highly universal coupling between knowledge and algorithms. Finally, simulation experiments are carried out in a 300×300 grid environment. Experimental results show that the proposed knowledge-enhanced algorithm significantly outperforms traditional algorithms in welding robot path planning, fully verifying the effectiveness of domain knowledge enhancement and the universality of the pluggable framework. This work provides an efficient and stable solution for path planning in complex welding scenarios.

The main contributions of this study are as follows:

- We combine Large Language Models with intelligent optimization algorithms for the first time to construct a knowledge-driven path planning framework, which solves the problems of insufficient domain knowledge and poor condition adaptability in traditional algorithms.
- We design a lightweight, pluggable knowledge enhancement module and adaptive cost function that require no changes to the original algorithm structure, enabling high generality and flexible adaptation.
- We validate the proposed method on three mainstream algorithms (ABC, GA, RRT*), providing an efficient and stable path planning scheme for complex welding scenarios.

II. Related Work

Welding robot path planning (WRPP) [4], [5], [6], [7] aims to generate an optimal welding path for robots under given constraints, with the objective of minimizing the time or distance required to complete welding tasks. As a key enabling technology for advanced manufacturing and welding engineering, its performance directly determines the stability and efficiency of the welding process. Early studies mainly focused on spot-welding tasks using a single welding robot [22], [23], in which the robot traverses all welding points sequentially and returns to the origin. This workflow is highly similar to the classic Traveling Salesman Problem (TSP) [24]. However, such modeling is only applicable to ideal spot-welding scenarios with fixed structures and simple constraints, and tends to result in path redundancy and slow convergence in complex unstructured welding environments.

With the rapid development of intelligent optimization algorithms, Marco Dorigo et al. proposed the Ant Colony Optimization (ACO) algorithm [25], which simulates the foraging behavior of ant colonies and achieves global optimization in the solution space through pheromone positive feedback. On this basis, Sharma et al. adopted classic path search algorithms such as Dijkstra's algorithm [26] to quickly obtain the shortest path using a greedy strategy. Strub et al. introduced an adaptive heuristic function into the A* algorithm [27], effectively improving path search efficiency in static structured environments. Wu et al. integrated the Genetic Algorithm (GA) [13] into the planning process to enhance global optimization capability via evolutionary operations, including crossover and mutation. Shen et al. applied the global search performance of the Artificial Bee Colony (ABC) [28] algorithm to welding robot path planning. Although traditional intelligent algorithms exhibit certain planning performance in simple static scenarios, they have obvious drawbacks in complex welding conditions, such as delayed pheromone update [29] and insufficient dynamic adaptability [30].

To address the above problems, researchers have attempted to improve the algorithms from a mechanism perspective. Wang et al. proposed an improved Ant Colony Optimization algorithm based on Monte Carlo (MC-IACO) [31], which replaces the traditional pheromone update mechanism with a Monte Carlo factor and effectively enhances the global exploration and local exploitation capabilities of the algorithm. Hasselt et al. presented an improved path planning algorithm based on Double DQN [32], [33], [34], achieving local performance optimization in specific complex scenarios. Tran et al. introduced a multi-sensor interaction strategy [35] to improve environmental perception accuracy. Rostami et al. proposed an improved artificial potential field obstacle avoidance algorithm [36], which exhibits satisfactory path planning and obstacle avoidance performance in a single static obstacle environment.

Despite certain progress achieved in the above studies, they still fail to deeply integrate cross-scenario knowledge in the welding domain. When facing complex and multi-source working environments, the algorithms exhibit weak adaptive matching capability and poor knowledge reusability, eventually resulting in excessively redundant planned paths.

In recent years, Large Language Models (LLMs) [21], [37] have offered novel solutions for WRPP tasks. LLMs possess strong knowledge extraction and semantic understanding capabilities [38], [39], [40], enabling the conversion of unstructured data such as text and images into standardized structured knowledge. This effectively compensates for the deficiencies in knowledge representation and adaptive decision-making of traditional methods. Han et al. fused environmental perception

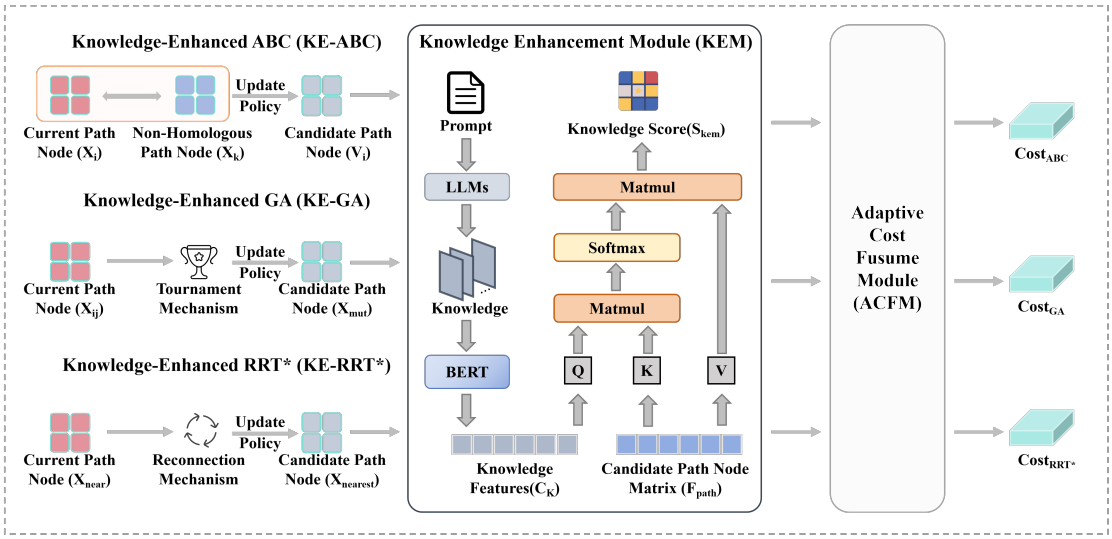


Fig. 2. Framework of the Domain Knowledge-Enhanced Welding Robot Path Planning Algorithm (DKE-WRPP).

knowledge with deep reinforcement learning [41], significantly enhancing the navigation adaptability of mobile robots in complex scenarios. Li et al. realized unified representation of multi-source heterogeneous information through large-scale model enhancement [42], providing crucial support for the application of cross-scenario navigation technology. Gómez-Espinosa et al. achieved high-precision extraction of weld paths based on depth RGB sensors [43], delivering accurate support for robot trajectory generation. Tran et al. fused visual perception features from multi-sensor fusion with prior welding operation knowledge, further improving the robustness of weld tracking and path planning [44].

In summary, existing methods exhibit obvious deficiencies in knowledge fusion and generalization. To this end, this study introduces welding domain knowledge into intelligent path planning optimization algorithms, and uses large language models to achieve domain knowledge extraction and fusion, effectively alleviating the poor generalization and insufficient dynamic adaptability of traditional methods. Meanwhile, the domain knowledge enhancement model constructed in this study is lightweight and pluggable, which can be flexibly embedded into mainstream path planning algorithms, significantly improving the stability and reliability of the algorithm in complex welding conditions.

III. Methods

A. Overview

The proposed Domain Knowledge-Enhanced Welding Robot Path Planning (DKE-WRPP) framework is shown in Fig. 2. Most previous path planning methods have obvious limitations: they fail to effectively mine and utilize welding-specific knowledge, and exhibit poor adaptability in complex welding scenarios, making it difficult to meet practical production requirements. To solve this problem, this study focuses on the core task of welding robot path planning and designs an improved framework with both knowledge-driven and condition-adaptive capabilities.

First, we employ the knowledge extraction capability of Large Language Models (LLMs) to accurately extract welding domain knowledge. Then, we use the pre-trained language model BERT to perform knowledge semantic encoding and generate standardized knowledge vectors. Next, we construct a Knowledge Enhancement Module (KEM) and introduce an attention mechanism to deeply fuse knowledge features and path geometric features. Finally, we update the cost functions of the three baseline algorithms through the Adaptive Cost Fusion Module (ACFM), realizing low-intrusive coupling between domain knowledge and path planning algorithms while ensuring the pluggability of each module in the framework to adapt to various welding path planning scenarios.

B. Simulation Environment Construction

To fully demonstrate the effectiveness and robustness of the DKE-WRPP framework, simulation experiments are conducted in a 300×300 grid environment. First, to simulate random obstacles such as spatters and fixtures in actual welding conditions, a random obstacle generation mechanism is introduced in the initialization phase. Static obstacles are randomly generated at different positions according to the obstacle distribution in a real welding workshop, providing an experimental environment close to actual working conditions for subsequent algorithm performance evaluation and ensuring the scientificity and reliability of experimental results.

First, we randomly generate K obstacles ($K \geq 100$) in the 2D workspace $W = \{(x, y) \mid x_{min} \leq x \leq x_{max}, y_{min} \leq y \leq y_{max}\}$ to simulate the distribution characteristics of obstacles such as spatters and fixtures in a real welding workshop. The set of obstacle center coordinates is denoted as $O_c = \{o_1, o_2, \ldots, o_K\}$. Then, the center coordinates (x, y) of obstacles are randomly sampled through the uniform distribution U:

$$x_k \sim U(x_{min}, x_{max}), \quad y_k \sim U(y_{min}, y_{max}), \tag{1}$$

where, the center coordinate of the k-th obstacle is denoted as $o_k = (x_k, y_k)$.

Considering the actual physical volume of obstacles and the safety boundary of the welding robot, we expand an expansion radius R_k outward from each random center point o_k to form a single obstacle entity area O_k. The total obstacle area O_{total} in the workspace can be expressed as the union of all random obstacle areas:

$$O_{total} = \bigcup_{k=1}^{K} \left\{ p \in W \;\middle|\; \|p - o_k\|_2 \leq R_k \right\} \tag{2}$$

where, p is any point in the motion trajectory $P = \{p_1, p_2, \ldots, p_M\}$ of the robot in the workspace, used to describe position coordinates in space; $\| \cdot \|_2$ denotes the Euclidean distance, used here to calculate the straight-line distance between point p and obstacle o_k.

Meanwhile, to ensure the reachability of path planning, a legitimacy check mechanism is applied when randomly generating obstacle points to automatically eliminate invalid obstacles that surround the start point p_{start} or target point p_{goal}. Finally, the valid random obstacle center set O is obtained:

$$O = O_{total} - O_n \tag{3}$$

$$O_n = \{o_k \mid \|p_{start} - o_k\|_2 < R_k \cup \|p_{goal} - o_k\|_2 < R_k\} \tag{4}$$

C. Knowledge Feature Representation

Welding trajectories in complex obstacle environments must satisfy both collision avoidance requirements and welding process specifications. Traditional algorithms struggle to effectively utilize process prior knowledge. This study uses Large Language Models to extract welding domain rules and generates high-dimensional semantic feature vectors via pre-trained BERT to provide computable prior inputs for the algorithm. Chain-of-Thought (CoT) [45] is adopted to design structured prompts $Prompt$, guiding the model to focus on environmental features, robot kinematic constraints, and welding process specifications to fully cover complex process requirements:

$$T = LLMs(Prompt), \tag{5}$$

where, $T = \{t_1, t_2, \ldots, t_N\}$ represents a structured set of domain knowledge, N is the total number of extracted domain rules, and t_i represents the $i-th$ specific knowledge description text, e.g., "The welding path should maintain stable weld bead direction continuity, prefer feasible, smooth, and target-close trajectories; reduce sharp turns and redundant detours while meeting process safety distance constraints, and avoid high-risk obstacle areas". These natural-language specification texts form the prior input for the subsequent knowledge enhancement module.

Then, deep semantic feature extraction is performed on the domain knowledge set T through the pre-trained language model $BERT$. For each knowledge text t_i in the set, we first discretize it into a token sequence using a tokenizer, and insert the token and token at the beginning and end of the sequence, respectively. The token sequence is then mapped into a dense continuous vector representation through the embedding layer:

$$E_i = \text{Embedding}(t_i) \in \mathbb{R}^{L_i \times d}, \tag{6}$$

where, Embedding$(\cdot)$ denotes the embedding layer mapping function of the BERT model, converting discrete token sequences into dense representations in a continuous vector space. Its output fuses Token Embedding and Position Embedding, encoding the semantic information of tokens and their positional information in the sequence, respectively. L_i is the length of the token sequence of the $i-th$ knowledge text after tokenization, and d is the embedding dimension.

Then, E_i is fed into the encoder of the Multi-Head Self-Attention mechanism to deeply parse the implicit logical connections between welding domain knowledge using its adaptive global context perception capability, obtaining a representation vector h_i containing global semantic information:

$$h_i = \text{BERT}(E_i) \in \mathbb{R}^d, \tag{7}$$

$$C_K = [h_1, h_2, \ldots, h_N]^T \in \mathbb{R}^{N \times d}, \tag{8}$$

where, d denotes the hidden layer dimension of the pre-trained language model. C_K is a standardized feature matrix (knowledge vector) containing all welding prior knowledge, which deeply integrates welding process semantic features and is completely decoupled from the underlying path planning algorithms. This vectorization method of domain prior knowledge provides theoretical support for the Knowledge Enhancement Module (KEM) to achieve high generality and pluggable coupling.

D. Knowledge Enhancement Module (KEM)

After obtaining the standardized knowledge vector C_K, this study constructs a Knowledge Enhancement Module (KEM). This module establishes a mapping between the high-dimensional semantic space and the geometric space through the cross-attention mechanism, evaluates the matching degree between the planned path and the prior knowledge of welding processes, and adaptively adjusts the cost function of the benchmark algorithm.

To ensure the generality of the framework, KEM is designed as a low-intrusion, plug-and-play module. It only guides the search direction through the cost function evaluation process without modifying the underlying path search algorithm.

Specifically, KEM first extracts features of path nodes and their adjacent poses from motion trajectory P to obtain the path feature matrix F_{path}. Then, it performs cross-attention calculation $CrossAtt(\cdot)$ between F_{path} and the knowledge vector C_K, realizing cross-modal fusion of geometric features and domain knowledge. Finally, the fused features are used to adjust the cost function, providing effective support for path optimization and evaluation.

$$CrossAtt\left(\mathcal{Q},\mathcal{K},\mathcal{V}\right)=softmax\left(\frac{\mathcal{Q}\mathcal{K}}{\sqrt{d_{attn}}}\right)\mathcal{V}, \quad (9)$$

$$S_{kem}(C_K,F_{path})=CrossAtt\left(C_K,F_{path},F_{path}\right), \quad (10)$$

where, the knowledge vector C_K acts as the query vector $\mathcal{Q}$, and the path feature matrix F_{path} is linearly projected into the key matrix $\mathcal{K}$ and value matrix $\mathcal{V}$. d_{attn} is the hidden layer dimension. $Softmax$ denotes the normalization function, which weights and sums the value matrix, finally outputting a knowledge matching score $S_{kem}(\cdot)$ ranging from 0 to 1. This score intuitively reflects the extent to which the current path follows the welding process priors defined in the text.

E. Adaptive Cost Fusion Module (ACFM)

To achieve seamless coupling with diverse baseline algorithms, this study designs a unified adaptive cost function update paradigm. For both global swarm intelligence algorithms (e.g., ABC, GA) for fitness evaluation and incremental sampling algorithms (e.g., RRT*) for local reconnection cost decision-making, the composite cost function $Cost_{total}(\cdot)$ is composed of the basic geometric cost $Cost_{geo}(\cdot)$ plus an adaptive knowledge penalty term:

$$Cost_{total}=Cost_{geo}(F_{path})+\lambda\cdot\left(1-S_{kem}(C_K,F_{path})\right), \quad (11)$$

where, λ is the knowledge gain coefficient, used to adjust the intervention weight of prior knowledge in the path optimization process. When the knowledge matching score S_{kem} approaches 1, the path highly complies with welding process specifications, and the knowledge penalty term tends to zero. Conversely, if the path has sharp turns or is close to dangerous obstacles, a high penalty cost is triggered. The pseudocode of the ACFM calculation process is shown in Algorithm 1.

Algorithm 1 Adaptive Cost Fusion Module (ACFM)

input: C_K: standardized knowledge vector;
$allowed_X$: list of accessible candidate nodes / paths;
$Cost_{geo}(\cdot)$: geometric cost function;
$S_{kem}(\cdot)$: cross-attention knowledge matching function;
λ: knowledge penalty gain coefficient.
output: Adaptive total cost array $Cost_{total}$.
Initialize $list_1, list_2 \leftarrow [\,]$. **for** $X_j \in allowed_X$ **do**:
$list_1[j] = \lambda \cdot (1 - S_{kem}(C_K, X_j))$
$list_2[j] = Cost_{geo}(X_j)$
end for
Fuse geometric cost with knowledge penalty:
$Cost_{total} \leftarrow list_2 + list_1$
return $Cost_{total}$

1) Knowledge-Enhanced ABC Algorithm (KE-ABC): This study proposes a Knowledge-Enhanced Artificial Bee Colony (KE-ABC) algorithm, which embeds the Knowledge Enhancement Module (KEM) into its core evolutionary mechanism to realize the joint drive of geometric features and welding prior knowledge, thereby avoiding the problem of traditional ABC algorithms [12] falling into local optima.

Specifically, in the employed bee phase, the algorithm first performs local mutation search in the neighborhood of the memorized path X_i, and its spatial position update strategy is formalized as:

$$V_i = X_i + \phi_i(X_i - X_k), \quad (12)$$

where, V_i represents the generated candidate path node, X_k is another non-homologous path node randomly selected from the population ($k \neq i$), and ϕ_i is a random disturbance factor uniformly distributed in $[-1, 1]$, used to control the step size of local exploration.

After generating the candidate path, we calculate the physical geometric cost $Cost_{geo}(V_i)$ of the candidate path node V_i and the knowledge matching score $S_{kem}(V_i, C_K)$ of the path with the knowledge vector C_K. Thus, the composite cost function $Cost_{ABC}$ fused with welding domain rules is defined as follows:

$$Cost_{ABC} = Cost_{geo}(V_i) + \lambda_{abc} \cdot (1 - S_{kem}(C_K, V_i)), \quad (13)$$

where, λ_{abc} is the knowledge gain weight coefficient.

Finally, in the scout bee phase, if the fitness $Cost_{ABC}$ of a path stagnates after multiple iterations, a reset mechanism is triggered.

2) Knowledge-Enhanced GA Algorithm (KE-GA): The Genetic Algorithm [13] (GA) relies on global parallel search and population evolution mechanisms. In multi-constrained welding tasks, it easily guides the population to evolve toward regions with infeasible spatial postures or drastic curvature changes, generating a large number of invalid iterations. To this end, this study introduces the cross-modal semantic evaluation of the Knowledge Enhancement Module (KEM) into the genetic algorithm to construct a Knowledge-enhanced Genetic Algorithm (KE-GA).

Specifically, KE-GA adopts Tournament Selection in the selection operator, screening individuals with the minimum composite fitness from the candidate subset with absolute probability to enter the mating pool. Then, uniform crossover is used to complete allele exchange of the local spatial topology of the parents. Finally, Gaussian random disturbance is introduced to apply physical position offset to the offspring waypoints, and its mutation update mechanism is formalized as:

$$X_{mut} = X_{ij} + \mathcal{N}(0, \sigma^2), \quad (14)$$

where, X_{mut} is the position of the $i-th$ individual in the j-dimensional space after mutation. $\mathcal{N}(0, \sigma^2)$ is Gaussian

noise with zero mean and variance σ^2. The introduction of Gaussian mutation aims to continuously expand the exploration boundary of the feature solution space and effectively prevent population homogenization from falling into local extremes.

Then, the algorithm reconstructs its fitness evaluation into a composite function $Cost_{GA}$ of basic geometric cost and knowledge penalty term, with the formula as follows:

$$Cost_{GA} = Cost_{geo}(X_{mut}) + \lambda_{ga} \cdot (1 - S_{kem}(C_K, X_{mut})), \tag{15}$$

where, λ_{ga} is the penalty weight regulating the strength of process prior constraints.

The new population after crossover and mutation operations is again re-evaluated for process compliance by the KEM module, supplemented by an elite retention strategy. KE-GA not only retains the breadth of global optimization but also continuously injects multi-modal welding prior knowledge into the genetic evolution process, significantly improving the industrial practicality of the final trajectory.

3) Knowledge-Enhanced RRT Algorithm (KE-RRT*):* The Optimal Rapidly-exploring Random Tree [14] (RRT*) algorithm relies solely on Euclidean distance metrics during tree expansion, easily generating trajectories close to obstacles. To this end, KE-RRT* deploys the Knowledge Enhancement Module (KEM) into the edge-level expansion logic.

Specifically, we calculate the knowledge matching score $S_{kem}(X_{nearest}, C_K)$ between the nearest node $X_{nearest}$ and the knowledge vector C_K to construct the cumulative composite cost $Cost_{RRT*}(X_{nearest})$ for local expansion:

$$\begin{aligned} Cost_{RRT*} = {} & Cost_{geo}(X_{nearest}) \\ & + \lambda_{rrt*} \cdot (1 - S_{kem}(C_K, X_{nearest})), \end{aligned} \tag{16}$$

where, λ_{rrt*} is the knowledge penalty gain coefficient.

To ensure both asymptotic optimality of the trajectory and process compliance, KE-RRT* reconstructs the core reconnection mechanism with knowledge. For any potential node X_{near} in the neighborhood, the judgment condition for the reconnection operation is defined as:

$$\begin{aligned} & Cost_{geo}(X_{near}) \\ & + \lambda_{rrt*} \cdot (1 - S_{kem}(C_K, X_{near})) < Cost_{RRT*}. \end{aligned} \tag{17}$$

Once the above formula holds, it indicates that the local path redirected via X_{near} has a better total length, and the algorithm immediately updates its parent node and cumulative cost.

IV. Experiments

In the following sections, systematic simulation experiments are conducted on the welding robot path planning task to comprehensively evaluate the proposed DKE-WRPP framework. This study takes three mainstream path planning algorithms as baselines, each connected to a pluggable Knowledge Enhancement Module (KEM). Simulation results show that DKE-WRPP significantly outperforms various baseline intelligent algorithms and verifies the generality of the pluggable knowledge framework.

A. Experimental Setup

1) System setup: In terms of hardware configuration, the experiment uses an NVIDIA A40 GPU as the core computing device, providing sufficient computing power for algorithm operation and domain knowledge encoding, effectively reducing the runtime of large-scale simulation experiments. In terms of software configuration, the experimental program is developed based on Python, and Matplotlib is used for visual display of results. All software adopts stable and compatible versions and has undergone multiple rounds of compatibility testing to ensure the stable operation of the entire experimental process.

TABLE I. Simulation Environment and Core Parameter Settings for Baseline Algorithms.

Module/Algorithm	Parameter Name	Set Value
Common Parameters	Grid space size	300×300
	Number of static obstacles	100
	Safety obstacle avoidance buffer / mm	5.0
	Knowledge module penalty weight (λ)	140.0
ABC Algorithm	Total colony size	30
	Maximum iterations	200
	Path control way points	8
	Stagnation tolerance limit	50
GA Algorithm	Population size	50
	Maximum iterations	200
	Crossover probability	0.8
	Mutation probability	0.15
	Elite retention ratio	10%
RRT* Algorithm	Maximum samples	4000
	Node expansion step size / mm	2.0
	Neighborhood reconnection radius / mm	10.0
	Goal-biased sampling probability	10%

2) Parameter settings: The parameter settings of the three baseline algorithms refer to typical configurations in the field of welding robot path planning [46], adjusted according to the 300×300 grid scenario and obstacle avoidance requirements of this experiment to ensure full performance of each algorithm, as shown in Table I. The number of obstacles is set to 100, the safety obstacle avoidance buffer distance is 5.0 mm, and the penalty weight of the knowledge module is set to 140.

a) Artificial Bee Colony Algorithm (ABC) (Table I): The total colony size is set to 30, including 15 employed bees and 15 onlooker bees, balancing the global exploration and local exploitation capabilities of the algorithm; the maximum number of iterations for the entire colony is set to 200; the maximum number of nodes allowed for the welding robot motion path is limited to 8; the stagnation tolerance limit is 50.

b) Genetic Algorithm (GA) (Table I): The population size is set to 50, the number of iterations is set to 200, the crossover probability is 0.8, and the mutation probability is 0.15, balancing population diversity and algorithm convergence speed; the selection operator adopts roulette wheel selection, the crossover operator adopts single-point crossover,

and the mutation operator adopts random mutation to improve algorithm optimization efficiency. Meanwhile, the elite retention ratio is 10%.

c) Optimal Rapidly-exploring Random Tree Algorithm (RRT) (Table I):* The sampling area is consistent with the 300×300 grid scenario; the maximum number of samples is set to 4000, the node expansion step size is 2 mm, and the neighborhood reconnection radius is 10 mm, ensuring the algorithm can quickly search for feasible paths; the goal-biased sampling probability is 10%, and the optimal solution is determined when the deviation between the sampling node and the end point coordinate is less than this threshold.

3) Evaluation metrics: To comprehensively evaluate the effect of the algorithm, four core evaluation metrics [47] are selected in the experiment, evaluated from four dimensions:

a) Path Length (unit: mm): The total length of the planned path from the start point to the end point of the robot.

b) Runtime (unit: s): The total time taken by the algorithm to complete one path planning.

c) Path Smoothness: The amplitude of angle change and curvature mutation between adjacent nodes in the planned path.

d) Iteration Convergence Count: For ABC and GA algorithms, it refers to the number of iterations to reach the optimal solution; for RRT* algorithm, it refers to the number of samples to reach the optimal solution.

TABLE II. COMPARISON OF PERFORMANCE METRICS FOR VARIOUS ALGORITHMS IN A 300 × 300 COMPLEX ENVIRONMENT.

Algorithm	Path Length(mm)	Smoothness	Iterations	Time(s)
ABC[12]	380.432	0.0334	214	1.15
GA[13]	382.105	0.0566	231	1.32
RRT*[14]	366.892	0.0487	190	0.78
KE-ABC	378.210	0.0215	142	1.98
KE-GA	377.554	0.0302	155	2.12
KE-RRT*	359.115	0.0240	124	1.85

B. Baselines

This experiment selects three mainstream intelligent algorithms widely used in the path planning field with stable performance as baselines. Their core logic and optimization mechanisms have different emphases, which can fully verify the generality of the Knowledge Enhancement Module (KEM).

1) Artificial Bee Colony (ABC) [12]: A swarm intelligence optimization algorithm simulating bee honey-collecting behavior. It balances global exploration and local exploitation through the collaboration of employed bees, onlooker bees, and scout bees, suitable for path optimization and obstacle avoidance tasks in continuous spaces.

2) Genetic Algorithm (GA) [13]: Based on biological evolution theory and natural selection mechanism, it iteratively optimizes through operations such as population initialization, crossover, mutation, and selection, with strong global optimization ability, which can effectively solve combinatorial optimization problems in path planning.

3) Optimal Rapidly-exploring Random Tree (RRT) [14]:* A sampling-based path planning algorithm. It gradually expands the tree structure through random sampling, which can quickly search for feasible paths and continuously optimize path quality, suitable for path planning in high-dimensional or complex obstacle scenarios.

C. Main Comparative Experiments

To verify the superiority and pluggability of the DKE-WRPP algorithm, experiments are conducted with the three algorithms ABC, GA, and RRT*, respectively, and performance comparisons are carried out around four core evaluation metrics. The experimental results are the average of 5 independent experiments to ensure the reliability of the results. The specific comparison results are shown in Table II. The three knowledge-enhanced algorithms (KE-ABC, KE-GA, KE-RRT*) connected to the pluggable KEM module are significantly superior to the corresponding baseline algorithms in the four core evaluation metrics, fully verifying the superiority of the DKE-WRPP algorithm and the effectiveness of the KEM module.

In terms of path length, KE-ABC, KE-GA, and KE-RRT* are shortened by 0.58%, 1.19%, and 2.12%, respectively compared with the baseline algorithms, indicating that the integration of domain knowledge can effectively guide the algorithm to search for paths closer to the optimal solution, reduce invalid detours, and improve path optimality. In terms of runtime, the knowledge-enhanced algorithms take slightly longer due to the introduction of knowledge feature fusion and adaptive cost calculation, but still remain within an acceptable range, proving that the low-intrusive coupling design of the KEM module does not cause excessive computational burden. In terms of path smoothness, KE-ABC, KE-GA, and KE-RRT* drop to 0.0215, 0.0302, and 0.0240 respectively, significantly lower than the original algorithms. The trajectory has smaller turning angles and more continuous curvature, more in line with the actual operation motion requirements of welding robots. In terms of iteration convergence count, KE-ABC, KE-GA, and KE-RRT* are reduced by 33.64%, 32.90%, and 34.74%, respectively compared with the baseline algorithms, fully reflecting the guiding role of domain knowledge in the algorithm optimization process, helping the algorithm quickly converge to the optimal solution and improving the optimization efficiency and stability of the algorithm.

To further intuitively show the optimization effect of the KEM module on algorithm convergence performance, this experiment plots the iteration convergence graphs of the three baseline algorithms and the corresponding knowledge-enhanced algorithms, as shown in Fig. 3.

Specifically, Fig. 3(a) shows the path planning comparison between the ABC algorithm and the KE-ABC algorithm. It can be seen from the figure that the KE-ABC path is shorter, with fewer detours and more regular node distribution, showing better performance. Meanwhile, Fig. 3(d) shows the convergence comparison between the two algorithms. It can be found that the ABC algorithm has a slow convergence speed in the early iteration (exploration phase). With the

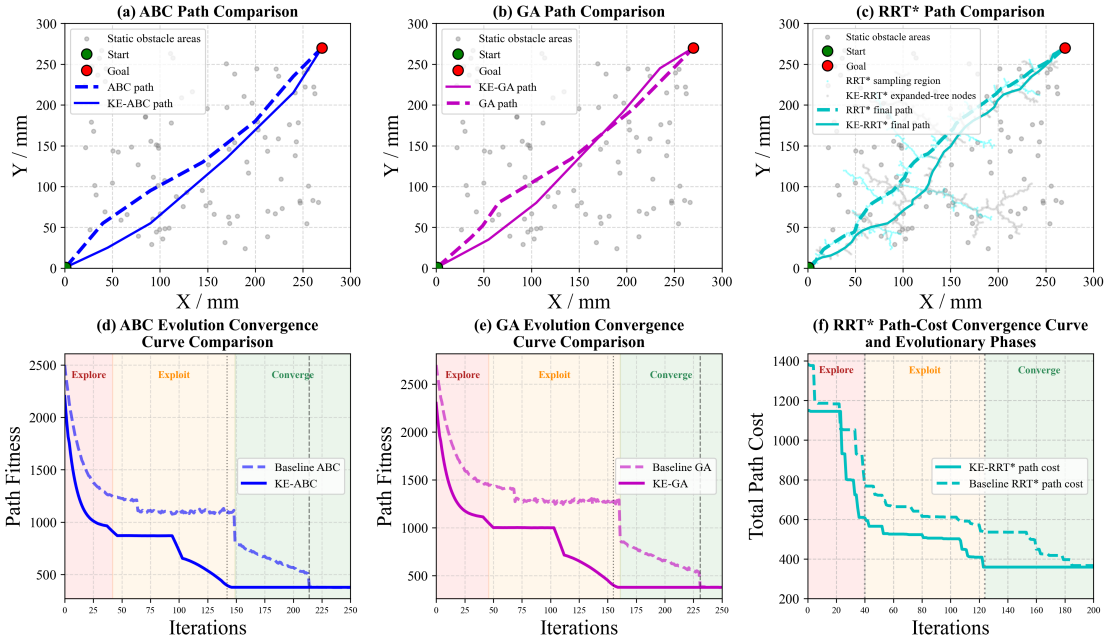


Fig. 3. Simulated path trajectories and convergence curve comparison between baseline and knowledge-enhanced algorithms.

increase of iteration times, the path length gradually decreases and stabilizes, finally reaching the optimal solution at about 214 iterations. At the same time, the KE-ABC algorithm has a significantly accelerated convergence speed, reaching the optimal solution at about 142 iterations, and the final converged path length is shorter, fully reflecting the optimization effect of the KEM module on the convergence performance of the ABC algorithm.

Similarly, Fig. 3(b) and Fig. 3(e) show the path planning comparison and iteration convergence graph between the GA algorithm and the KE-GA algorithm. The global optimization process of the GA algorithm is relatively lengthy, reaching the optimal solution at about 231 iterations. Under the guidance of knowledge, the iteration count of KE-GA drops to 155, with a smoother path, shorter length, faster decline of the convergence curve and lower stable value, effectively overcoming the shortcomings of traditional GA such as easy prematurity and poor path smoothness.

Secondly, Fig. 3(c) and Fig. 3(f) show the expansion tree comparison of the RRT* algorithm. Since RRT* is a tree-based search algorithm based on sampling, the path is generated by connecting expansion tree nodes, so the expansion tree is used for comparison. Under knowledge constraints, KE-RRT* has fewer sampling nodes, a more compact tree structure, away from high-risk obstacle areas, smoother reconnection paths, and the convergence sampling count drops from 190 to 124. While maintaining the fast search advantage of RRT*, it significantly improves path optimality and process compliance.

In summary, KE-ABC, KE-GA, and KE-RRT* algorithms outperform the traditional baseline methods significantly in path length, smoothness, and convergence speed, with the number of convergence iterations reduced by more than 33% on average. The planned paths are shorter, smoother, and more consistent with welding process constraints. Experimental results fully demonstrate that the proposed pluggable domain knowledge enhancement framework can effectively improve the planning performance of traditional intelligent optimization algorithms in complex welding scenarios with low intrusion and high generality, showing strong practicality and generalization ability.

TABLE III. PATH PLANNING PERFORMANCE COMPARISON USING DIFFERENT MULTIMODAL LARGE LANGUAGE MODELS IN DKE-WRPP.

Model	Path length (mm)	Smoothness	Iterations	Time(s)
KE-ABC	378.210	0.0215	142	1.98
GPT-4O-mini[48]	378.210	0.0215	142	1.98
Qwen2.5-72B[49]	378.91	0.0223	148	3.99
Qwen2.5-7B[49]	380.12	0.0238	155	2.01
InternVL2.5-8B[50]	381.34	0.0252	163	2.03
LLaVA1.5-7B[51]	382.56	0.0267	171	2.05

D. Knowledge Accuracy Impact Experiment

To test the relationship between the optimization effect of the domain knowledge enhancement module and the generation ability of large language models, this experiment selects five mainstream Large Language Models for knowledge accuracy comparison experiments: Qwen2.5-7B [49], Qwen2.5-72B [49], GPT-4o-mini [48], InternVL2.5-8B [50], and LLaVA1.5-7B [51]. At the same time, the KE-ABC algorithm is used as the benchmark under the same experimental environment and parameter settings, and performance comparisons are carried out around the four evaluation metrics set above. The experimental results are the average of 5 independent experiments to measure the effect of knowledge encoding generated by different large models. The specific experimental results are shown in Table III.

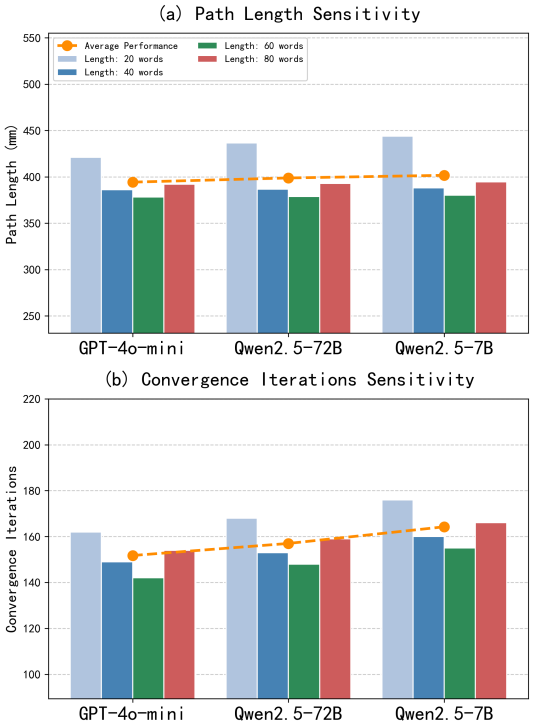


Fig. 4. Sensitivity of path planning performance to knowledge length.

The experimental results show that the knowledge encoding generated by different Large Language Models has significant differences in the optimization effect of the KE-ABC algorithm, and the detail level of knowledge generated by the model has an obvious positive correlation with the algorithm optimization effect. Specifically, the optimization effect ranking of each model is completely consistent with the knowledge detail level ranking, followed by GPT-4o-mini > Qwen2.5-72B > Qwen2.5-7B > InternVL2.5-8B > LLaVA1.5-7B.

Among them, the welding domain knowledge generated by GPT-4o-mini is the most complete and accurate, and the corresponding KE-ABC algorithm has the best performance, with a path length of 378.210 mm, path smoothness of 0.0215, convergence iteration count of 142, and runtime of 1.98 s, which is completely consistent with the main experimental results of this study. The knowledge quality generated by Qwen2.5-72B is slightly lower than that of GPT-4o-mini, and each indicator follows closely, with overall performance better than Qwen2.5-7B with smaller parameters. InternVL2.5-8B and LLaVA1.5-7B are weak in understanding welding process constraints and completeness of knowledge extraction, resulting in increased planned path length, decreased trajectory smoothness, and increased convergence iteration counts. Among them, LLaVA1.5-7B has the worst performance due to sparse knowledge and insufficient coverage of key constraints.

In summary, the experimental results clearly show that the higher the knowledge generation quality of the Large Language Model, the more accurate the guidance to the DKE-WRPP algorithm optimization process, which can effectively shorten the path length, improve trajectory smoothness, and accelerate convergence speed. This conclusion fully verifies the robustness and effectiveness of the KEM module, and provides an objective basis for Large Language Model selection and welding domain knowledge generation strategies in practical engineering applications.

E. Feasibility Analysis and Experimental Design of LLM Knowledge Length Experiment

To verify the impact of the length of knowledge encoding generated by large language models on the optimization effect of the KE-ABC algorithm, this experiment strictly follows the hardware environment, software configuration, algorithm parameters, and four evaluation metrics of the previous knowledge accuracy experiment, and only takes the knowledge encoding length as the only variable to conduct a controlled experiment. The experiment selects the three best-performing multimodal large models (GPT-4o-mini, Qwen2.5-72B, Qwen2.5-7B) for testing, and the results are shown in Fig. 4.

As shown in the path length sensitivity curve [Fig. 4(a)] and the convergence iteration sensitivity curve [Fig. 4(b)], the length of knowledge encoding is significantly correlated with the optimization performance of the algorithm, and there exists an obvious optimal length interval. The KE-ABC algorithm achieves the best performance when the knowledge length is set to 60 words. In this case, the path length is minimized, the number of convergence iterations is the lowest, and the core domain knowledge such as welding process constraints and trajectory smoothness can be fully covered without redundant information interference, so that the guidance of knowledge is fully exerted.

When the knowledge encoding length is 20 words, the excessively short length results in the loss of key process constraints and incomplete knowledge representation, which weakens the guidance for algorithm optimization. Consequently, the path length increases significantly, the convergence iterations rise substantially, and the overall performance degrades sharply. Similarly, when the knowledge encoding length reaches 80 words, redundant semantic information increases the computational overhead of knowledge encoding and cross-attention. The guidance gain for the algorithm becomes saturated, and some indicators even decline slightly.

In terms of model differences, large-parameter models such as GPT-4o-mini and Qwen2.5-72B show stronger robustness to changes in knowledge length and a wider optimal interval. In contrast, Qwen2.5-7B is more sensitive to length, and its performance drops more significantly when deviating from the optimal interval.

In summary, this study determines that the optimal length for welding domain knowledge encoding is 60 words. The knowledge length experiment further verifies the effectiveness and robustness of the proposed knowledge enhancement framework, identifies the optimal balance between the scale and integrity of domain knowledge, and provides a reliable basis for the standardized design of knowledge encoding in practical engineering applications.

V. Conclusion

To address the limitations of existing welding robot path planning (WRPP) algorithms, particularly the insufficient in-

tegration of welding domain knowledge and poor adaptability to complex working conditions, this study proposes the DKE-WRPP algorithm based on domain knowledge enhancement along with a plug-and-play knowledge fusion framework. Domain-specific knowledge of the welding process is extracted via Large Language Models (LLMs) and semantically encoded to construct a Knowledge Enhancement Module (KEM); an attention mechanism is then employed to deeply fuse domain knowledge with path features, enabling the adaptive optimization of three conventional intelligent algorithms: ABC, GA, and RRT*. Systematic experiments on all three algorithms fully validate the effectiveness of domain knowledge enhancement and the generalizability of the plug-and-play framework. This work effectively addresses the core limitations of existing WRPP algorithms, offering a practical and viable solution for welding automation, with significant implications for advancing the technological development of welding robot path planning. Future work will focus on the deep integration of lightweight knowledge agents with welding robots, continuously optimizing the adaptive capability and execution efficiency of path planning, thereby providing theoretical and technical support for the realization of highly intelligent welding engineering.

ACKNOWLEDGMENT

This work was supported by The Liaoning Provincial Science and Technology Plan Joint Program (General Project of Natural Science Foundation) (JH4/4800, No.2025-MSLH-012).

REFERENCES

[1] R. G. Prasuna and S. R. Potturu, "Deep reinforcement learning in mobile robotics–a concise review," *Multimedia Tools and Applications*, vol. 83, no. 28, pp. 70 815–70 836, 2024.

[2] B. Siciliano and O. Khatib, "Robotics and the handbook," in *Springer Handbook of Robotics*. Springer, 2016, pp. 1–6.

[3] L. Chen, Z. Jiang, L. Cheng, A. C. Knoll, and M. Zhou, "Deep reinforcement learning based trajectory planning under uncertain constraints," *Frontiers in Neurorobotics*, vol. 16, p. 883562, 2022.

[4] X. Wang, X. Zhou, Z. Xia, and X. Gu, "A survey of welding robot intelligent path optimization," *Journal of Manufacturing Processes*, vol. 63, pp. 14–23, 2021.

[5] J. Zhong, T. Wang, and L. Cheng, "Collision-free path planning for welding manipulator via hybrid algorithm of deep reinforcement learning and inverse kinematics," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 1899–1912, 2022.

[6] Y. Yan, P. Zhang, C. Du, H. Wang, H. Wang, D. Pi, and Y.-H. Chen, "Vehicle trajectory prediction for autonomous driving applications: State-of-the-art review, research challenges, and future directions," *Automotive Innovation*, vol. 9, pp. 263–295, 2026.

[7] N. A. Madjid, A. Ahmad, M. Mebrahtu, Y. Babaa, A. Nasser, S. Malik, B. Hassan, N. Werghi, J. Dias, and M. Khonji, "Trajectory prediction for autonomous driving: Progress, limitations, and future directions," *Information Fusion*, vol. 126, p. 103588, 2026.

[8] X. Wang, J. Wei, X. Zhou, Z. Xia, and X. Gu, "Dual-objective collision-free path optimization of arc welding robot," *IEEE robotics and automation letters*, vol. 6, no. 4, pp. 6353–6360, 2021.

[9] J. Xin, C. Meng, A. D'Ariano, F. Schulte, J. Peng, and R. R. Negenborn, "Energy-efficient routing of a multirobot station: A flexible time-space network approach," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 3, 2022.

[10] L. Guo and H. Zhang, "Autonomous mobile welding robot for discontinuous weld seam recognition and tracking," *The International Journal of Advanced Manufacturing Technology*, vol. 119, no. 7, pp. 5497–5509, 2022.

[11] G.-H. Song, C.-M. Lee, and D.-H. Kim, "Investigation of path planning to reduce height errors of intersection parts in wire-arc additive manufacturing," *Materials*, vol. 14, no. 21, p. 6477, 2021.

[12] D. Karaboga *et al.*, "An idea based on honey bee swarm for numerical optimization," 2005.

[13] L. Wu, K. Cui, and S.-B. Chen, "Redundancy coordination of multiple robotic devices for welding through genetic algorithm," *Robotica*, vol. 18, no. 6, pp. 669–676, 2000.

[14] X. Zhou, X. Wang, and X. Gu, "An approach for solving the three-objective arc welding robot path planning problem," *Engineering Optimization*, vol. 55, no. 4, pp. 650–667, 2023.

[15] S. Bhagat, H. Banerjee, Z. T. Ho Tse, and H. Ren, "Deep reinforcement learning for soft, flexible robots: Brief review with impending challenges," *Robotics*, vol. 8, no. 1, p. 4, 2019.

[16] H. Sun, W. Zhang, R. Yu, and Y. Zhang, "Motion planning for mobile robots—focusing on deep reinforcement learning: A systematic review," *IEEE Access*, vol. 9, pp. 69 061–69 081, 2021.

[17] A. Devo, G. Mezzetti, G. Costante, M. L. Fravolini, and P. Valigi, "Towards generalization in target-driven visual navigation by using deep reinforcement learning," *IEEE Transactions on Robotics*, vol. 36, no. 5, pp. 1546–1561, 2020.

[18] F. Quiroga, G. Hermosilla, G. Farias, E. Fabregas, and G. Montenegro, "Position control of a mobile robot through deep reinforcement learning," *Applied Sciences*, vol. 12, no. 14, p. 7194, 2022.

[19] B. Sangiovanni, G. P. Incremona, M. Piastra, and A. Ferrara, "Self-configuring robot path planning with obstacle avoidance via deep reinforcement learning," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 397–402, 2020.

[20] A. Candra, M. A. Budiman, and K. Hartanto, "Dijkstra's and a-star in finding the shortest path: A tutorial," in *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*. IEEE, 2020, pp. 28–32.

[21] S. Meng, Y. Wang, C.-F. Yang, N. Peng, and K.-W. Chang, "Llm-a*: Large language model enhanced incremental heuristic search on path planning," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 1087–1102.

[22] Y.-T. Zhao, L. Gan, W.-G. Li, and A. Liu, "Path planning of spot welding robot based on multi-objective grey wolf algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 6, pp. 6181–6189, 2021.

[23] Q. Zhang and M.-Y. Zhao, "Minimum time path planning of robotic manipulator in drilling/spot welding tasks," *Journal of Computational Design and Engineering*, vol. 3, no. 2, pp. 132–139, 2016.

[24] S. Lin and B. W. Kernighan, "An effective heuristic algorithm for the traveling-salesman problem," *Operations research*, vol. 21, no. 2, pp. 498–516, 1973.

[25] M. Dorigo, "The any system optimization by a colony of cooperating agents," *IEEE Trans. System, Man & Cybernetics-Part B*, vol. 26, no. 1, pp. 1–13, 1996.

[26] G. Sharma, S. Jain, and R. S. Sharma, "Path planning for fully autonomous uavs-a taxonomic review and future perspectives," *IEEE Access*, vol. 13, pp. 13 356–13 379, 2025.

[27] M. P. Strub and J. D. Gammell, "Adaptively informed trees (ait*): Fast asymptotically optimal path planning through adaptive heuristics," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3191–3198.

[28] Y. Shen, Y. Gao, M. Yuan, H. Sun, and Z. Guo, "Multi-objective immune optimization of path planning for ship welding robot," *Electronics*, vol. 12, no. 9, p. 2040, 2023.

[29] K. Hu, Q. Dong, Z. Ma, B. Li, and T. Qi, "Welding robot path planning based on ant colony algorithm," in *2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*. IEEE, 2024, pp. 13–18.

[30] Y. Luo, S. Lin, Y. Wang, and K. Liang, "Path planning for mobile robots in complex environments based on enhanced sparrow search algorithm and dynamic window approach," *Robotica*, vol. 43, no. 5, pp. 1929–1952, 2025.

[31] T. Wang, L. Wang, D. Li, J. Cai, and Y. Wang, "Monte carlo-based improved ant colony optimization for path planning of welding robot," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 7, p. 101603, 2023.

[32] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[33] S. Fang, "Reinforcement learning-driven deep learning approaches for optimized robot trajectory planning," *Scientific Reports*, vol. 15, no. 1, p. 37898, 2025.

[34] B. Zhao, Y. Wu, C. Wu, and R. Sun, "Deep reinforcement learning trajectory planning for robotic manipulator based on simulation-efficient training," *Scientific Reports*, vol. 15, no. 1, p. 8286, 2025.

[35] C.-C. Tran and C.-Y. Lin, "An intelligent path planning of welding robot based on multisensor interaction," *IEEE sensors journal*, vol. 23, no. 8, pp. 8591–8604, 2023.

[36] S. M. H. Rostami, A. K. Sangaiah, J. Wang, and X. Liu, "Obstacle avoidance of mobile robots using modified artificial potential field algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 1–19, 2019.

[37] C. Cui, Y. Ma, S.-Y. Park, Z. Yang, Y. Zhou, P. Liu, J. Lu, J. Peng, J. Zhang, R. Zhang, L. Li, Y. Chen, J. H. Panchal, A. Abdelraouf, R. Gupta, K. Han, and Z. Wang, "LLM4AD: Large language models for autonomous driving—concept, review, benchmark, experiments, and future trends," *Proceedings of the IEEE*, pp. 1–36, 2026.

[38] G. Zhang, W. Li, Q. Qian, J. Wang, D. Zhao, S. Lu, and R. Xu, "On the generalization capacities of mllms for spatial intelligence," *arXiv preprint arXiv:2603.06704*, 2026.

[39] H. Zhao, Z. Guo, C. Zhang, Z. Zhou, and Z. Hu, "Structuthink: Reasoning with task transition knowledge for autonomous llm-based agents," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 24 489–24 506.

[40] Z. Xu, Z. Wang, X. Jiang, D. Li, D. Cheng, and N. Wang, "Reasoning-driven multimodal llm for domain generalization," *arXiv preprint arXiv:2602.23777*, 2026.

[41] R. Han, S. Chen, S. Wang, Z. Zhang, R. Gao, Q. Hao, and J. Pan, "Reinforcement learned distributed multi-robot navigation with reciprocal velocity obstacle shaped rewards," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 5896–5903, 2022.

[42] Z. Li, X. Meng, X. He, Y. Zhang, and W. Yin, "Large-scale model-enhanced vision-language navigation: Recent advances, practical applications, and future challenges," 2026.

[43] A. Gómez-Espinosa, J. B. Rodríguez-Suárez, E. Cuan-Urquizo, J. A. E. Cabello, and R. L. Swenson, "Colored 3d path extraction based on depth-rgb sensor for welding robot trajectory generation," *Automation*, vol. 2, no. 4, pp. 252–265, 2021.

[44] C.-C. Tran and C.-Y. Lin, "An intelligent path planning of welding robot based on multisensor interaction," *IEEE sensors journal*, vol. 23, no. 8, pp. 8591–8604, 2023.

[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[46] C. Duan and S. Luo, "Research on welding robot path planning based on genetic algorithm," in *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*. IEEE, 2023, pp. 1401–1407.

[47] M. Dobiš, M. Dekan, P. Beňo, F. Duchoň, and A. Babinec, "Evaluation criteria for trajectories of robotic arms," *Robotics*, vol. 11, no. 1, p. 29, 2022.

[48] D. B. Dangi, B. B. Dangi, and O. Steinbock, "Evaluation of gpt-4o and gpt-4o-mini's vision capabilities for compositional analysis from dried solution drops," *ACS omega*, vol. 10, no. 18, pp. 18 955–18 959, 2025.

[49] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.

[50] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[51] X. An, Y. Xie, K. Yang, W. Zhang, X. Zhao, Z. Cheng, Y. Wang, S. Xu, C. Chen, D. Zhu *et al.*, "Llava-onevision-1.5: Fully open framework for democratized multimodal training," *arXiv preprint arXiv:2509.23661*, 2025.

Optimizing Whiteboard Digitization with Clarix: An Automated Vision-Based System for Real-Time Text Extraction and Surface Cleaning

Abdellah NABOU, Ameksa Mohammed, Ezzahoud Hajar, Bazgour Yassine, Abouhane Zahra
Faculty of Science Semlalia (FSSM), Cadi Ayyad University (UCA), LISI Laboratory, Marrakech, Morocco

***Abstract*—Whiteboard digitization in educational settings grapples with challenges posed by handwritten text, mathematical notation, and unstructured layouts that confound traditional Optical Character Recognition (OCR) systems. Clarix, a low-cost embedded system, addresses these issues by integrating a Raspberry Pi, high-resolution camera, and servo-driven erasure mechanism to automate content capture, text extraction, and archiving as searchable PDFs. This study benchmarks five multimodal Large Language Models (LLMs) (GPT-4o, Claude, Gemini, DeepSeek, Grok) against traditional OCR systems (PaddleOCR, EasyOCR, TesseractOCR) for extracting whiteboard content across mathematics, physics, and economics domains. Results highlight a stark performance divide: multimodal LLMs achieved F1-scores ranging from 0.7550 to 0.8466, with Gemini leading at 0.8466 (precision 0.8188, recall 0.8784), followed by GPT-4o (F1=0.8162) and DeepSeek (F1=0.7965), while PaddleOCR topped traditional systems with an F1-score of 0.3333 (precision 0.3523, recall 0.3223), followed by EasyOCR (0.1158) and TesseractOCR (0.0000). Notably, increased region detection correlated with diminished performance, underscoring the superiority of contextual understanding over exhaustive segmentation. Clarix's fusion of intelligent automation and advanced text processing marks a transformative advancement in bridging analog and digital educational environments.**

***Keywords*—Whiteboard digitization; multimodal Large Language Models (LLMs); Optical Character Recognition**

I. Introduction

A. Problematics

Traditional whiteboards remain prevalent in educational and collaborative environments due to their simplicity, affordability, and ease of use [1]. However, they lack modern capabilities such as digital archiving, real-time content sharing, and efficient management of written material [2]. Manual transcription of whiteboard content is time-consuming and error-prone, while frequent cleaning disrupts the workflow and reduces productivity [3]. Although smart boards and interactive digital whiteboards offer advanced features, their high cost and infrastructural demands hinder widespread adoption, especially in resource-constrained settings [4]. This creates a clear need for an affordable, automated solution capable of capturing, digitizing, and erasing whiteboard content in real-time—bridging the gap between traditional writing surfaces and modern digital tools [5].

B. Aim

The primary aim of this study is to develop a low-cost, intelligent whiteboard automation system that integrates embedded systems with advanced artificial intelligence to enhance traditional educational environments. The system seeks to automatically capture, process, and digitize whiteboard content while providing seamless erasure functionality to maintain continuous workflow [6].

The research focuses on leveraging computer vision and machine learning technologies to extract and structure textual information from whiteboard images, creating searchable digital archives that preserve educational content for future reference [7]. By combining automated content capture with intelligent text processing, the system aims to transform traditional whiteboards into smart educational tools without requiring significant infrastructure changes or investment [8].

This approach addresses the growing need for digital integration in educational settings, while maintaining the familiar and intuitive nature of conventional whiteboards. The resulting system is designed to facilitate enhanced accessibility, content reuse, and seamless integration between traditional teaching methods and modern digital documentation requirements [9].

C. Contributions of this Work

Traditional OCR systems remain limited when dealing with handwritten whiteboard content, mathematical expressions, irregular layouts, and varying lighting conditions commonly encountered in real educational environments [13]. Although commercial smart boards provide advanced digital functionalities, their high acquisition and maintenance costs restrict their adoption, particularly in resource-constrained institutions. Furthermore, most existing automated whiteboard systems focus either on mechanical cleaning or on text digitization independently, without integrating both functionalities into a unified intelligent embedded platform. Consequently, there is a growing need for an affordable, intelligent, and automated solution capable of simultaneously capturing, understanding, archiving, and cleaning whiteboard content in real time while preserving the simplicity and accessibility of traditional whiteboards.

The remainder of this study is organized as follows: Section II reviews related work on automated whiteboard systems, OCR technologies, and multimodal language models. Section III presents the proposed methodology, including the embedded system architecture, dataset preparation, evaluated models, and performance metrics. Section IV discusses the experimental results obtained from both traditional OCR systems and multimodal LLMs. Section V provides an analysis and discussion of the findings, including the strengths and

limitations of the proposed approach. Finally, Section VI concludes the study and outlines potential directions for future work.

II. Related Work

In recent years, the automation of whiteboard and chalkboard management has garnered attention due to its potential to enhance classroom and collaborative efficiency. A range of efforts have explored different facets of this problem, from mechanical board erasing systems to vision-based scribble detection algorithms. However, most existing solutions address these challenges in isolation—focusing either on physical cleaning mechanisms or on content recognition—but rarely both.

A notable mechanical contribution comes from the work titled "Automatic Whiteboard Cleaner Using Arduino and IR Sensor", which implements an Arduino-controlled mechanism that detects the presence of a user and automatically moves a brush across the whiteboard using a motorized assembly [24]. While it offers a practical and low-cost erasure mechanism, it is strictly mechanical in nature and does not address any form of content preservation or digitization. This typifies a broader class of embedded solutions that prioritize automation of the erasing process without incorporating any form of content awareness or digital archiving.

Similarly, earlier work by [19] proposed a vision-based chalkboard eraser that utilizes Hough Transform techniques to detect chalk scribbles and perform selective cleaning. Although innovative in its partial cleaning approach, the system lacked the ability to extract semantic content, making it unsuitable for educational scenarios that require content archiving. Other hardware-oriented efforts, such as PIC-controlled electronic dusters and Arduino-based remote erasers, have explored various motor control mechanisms—linear actuators, pulley systems, or servo-motor configurations—to enhance the speed and accuracy of board cleaning. Yet these systems remain limited to the mechanical domain and do not engage with the board content in an intelligent manner.

In parallel, advancements in Optical Character Recognition (OCR) have significantly improved the accuracy of text extraction from complex surfaces, including whiteboards. Modern OCR systems—such as Google's Tesseract OCR [11], Microsoft's Read API [27], and domain-specific models trained on handwriting or smudged whiteboard content [25]—have enabled high-quality digitization of visual information. These systems leverage deep learning techniques, convolutional neural networks (CNNs), and attention mechanisms to improve recognition in noisy, low-contrast, or non-standard fonts. However, these OCR systems are typically applied in post-processing scenarios and are not often integrated into real-time embedded systems.

Moreover, while some commercial smart boards offer digital archiving and touchscreen-based input [20], they come with prohibitive costs and require specialized infrastructure. Their reliance on proprietary ecosystems also limits adaptability and scalability, particularly in resource-constrained educational environments. Importantly, such solutions do not address the problem for traditional whiteboards still in widespread use across classrooms, laboratories, and meeting spaces.

Despite recent advances in OCR and smart educational technologies, existing solutions still suffer from major limitations when applied to real-world whiteboard environments. Traditional OCR systems struggle with handwritten content, mathematical symbols, and irregular layouts, while commercial smart boards remain expensive and inaccessible in many educational settings. Furthermore, current automated whiteboard cleaning systems rarely integrate intelligent content digitization and semantic extraction within a unified embedded platform. Therefore, there is a need for an affordable and intelligent solution capable of simultaneously capturing, understanding, archiving, and cleaning whiteboard content in real time.

The proposed Clarix system aims to fill this void by introducing a novel robotic prototype that unifies these two traditionally separate domains. It features a servo-driven mechanical brush system for automated whiteboard erasing and a camera module for capturing the board content once full. The captured image is processed using advanced OCR techniques, and the text is saved as a searchable PDF, enabling long-term storage, accessibility, and knowledge reuse. Additionally, a benchmarking study is conducted using state-of-the-art LLMs (such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini) to assess their effectiveness in real-world whiteboard digitization scenarios. This integrated approach represents a first-of-its-kind contribution to the field of intelligent educational tools, offering an affordable, scalable, and intelligent solution for analog-to-digital whiteboard transformation.

III. Methodology

The Clarix system evaluation employed a comprehensive methodology encompassing three core components: hardware implementation, data collection, and computational analysis. To validate the system's effectiveness, a custom embedded prototype was developed to demonstrate the physical automation capabilities, while a specialized dataset of whiteboard images was created to benchmark text extraction performance. The extraction capabilities were then evaluated using both traditional OCR engines and modern LLM-based approaches to determine the optimal digitization strategy for real-world deployment.

A. Embedded System

To physically implement the Clarix system's automation capabilities, a custom-designed embedded hardware prototype was developed. The embedded system is responsible for orchestrating the entire digitization and erasure cycle—from capturing the whiteboard content to initiating the motorized cleaning process. Designed to be cost-effective and easily replicable, the system is built using widely available and low-power electronic components, ensuring accessibility and scalability for educational and collaborative environments.

The overall hardware architecture of the proposed Clarix system is illustrated in Fig. 1, showing the integration of the Raspberry Pi controller, camera module, servo-driven cleaning mechanism, and peripheral components

At the core of the system is a Raspberry Pi 4, which serves as the primary computational unit. This microcomputer handles peripheral interactions, control signals, and system

logic. A high-resolution camera module, interfaced with the Raspberry Pi, is mounted at the top of the whiteboard frame to capture images of the content once the board is filled with text, equations, or diagrams. These images are then processed either locally or forwarded to external LLM APIs for intelligent text extraction and archiving, depending on future deployment choices.

The physical cleaning mechanism consists of a servo motor-driven brush mounted on a gear rack system, which allows smooth horizontal movement across the width of the whiteboard. The gear rack ensures precise control and linear motion, enabling a uniform and effective single-pass erasure. The motor is connected to a lightweight brush that spans the width of the board and is guided by wooden side supports, ensuring stability and alignment during motion.

Control of the brush is automated through the Raspberry Pi's GPIO pins, which trigger the servo motor upon completion of the image capture and text extraction steps. To provide real-time user feedback, the system is equipped with LED indicators that signal the system's current state—image capture (yellow), extraction in progress (blue), and cleaning in progress or completed (green). A buzzer serves as an auditory indicator that alerts the user before and after the cleaning process to ensure no disruption or accidental interference occurs.

The system also includes a physical power switch to ensure safe startup and shutdown, along with a manual button that acts as a trigger to initiate the full digitization-cleaning sequence. All connections are made via jumper wires and resistors, integrated on a modular breadboard circuit to allow easy testing, debugging, and future extensions.

The overall prototype is mounted on a compact frame designed to fit a standard-sized whiteboard and is entirely self-contained, with no dependency on external computing systems during operation. This embedded solution bridges the gap between mechanical automation and digital intelligence, offering an affordable yet advanced system for real-time whiteboard content management. To improve reproducibility and provide clearer visualization of the proposed architecture, a photograph of the fully assembled Clarix prototype is presented alongside the system diagram. The prototype integrates all hardware components within a compact self-contained structure, including the Raspberry Pi controller, camera module, servo-driven cleaning mechanism, status indicators, and supporting circuitry.

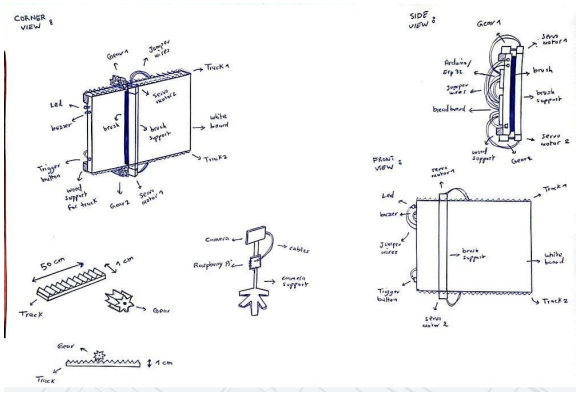


Fig. 1. Clarix prototype sketch showing the Raspberry Pi, servo motor, camera, and peripheral components for automated whiteboard cleaning.

B. Data

To evaluate the performance of both traditional OCR engines and LLM-based APIs in extracting whiteboard content, a custom dataset was created specifically for this study. The dataset consists of high-resolution images captured from a physical whiteboard where various types of content were manually written, including lecture notes, mathematical equations, tabular data, and bullet points. These samples were designed to simulate real-world classroom and meeting scenarios, incorporating common challenges such as variable handwriting styles, lighting reflections, marker inconsistencies, and non-uniform layouts. The images were taken using the camera module of the Clarix prototype under natural and artificial lighting conditions to ensure diversity in input. This dataset served as a benchmark for testing the accuracy and reliability of different text extraction techniques.

Prior to text extraction, all images were subjected to a standardized preprocessing pipeline specifically designed for traditional OCR engines to optimize their performance on handwritten whiteboard content. This preprocessing included image resizing for computational efficiency, Non-Local Means Denoising for noise reduction, and Contrast Limited Adaptive Histogram Equalization (CLAHE) for contrast enhancement. Notably, these preprocessing steps were applied exclusively to traditional OCR systems and not to LLM-based APIs, allowing for evaluation of the latter's native robustness in handling unprocessed whiteboard images.

A representative sample image from the dataset is shown in Fig. 2.

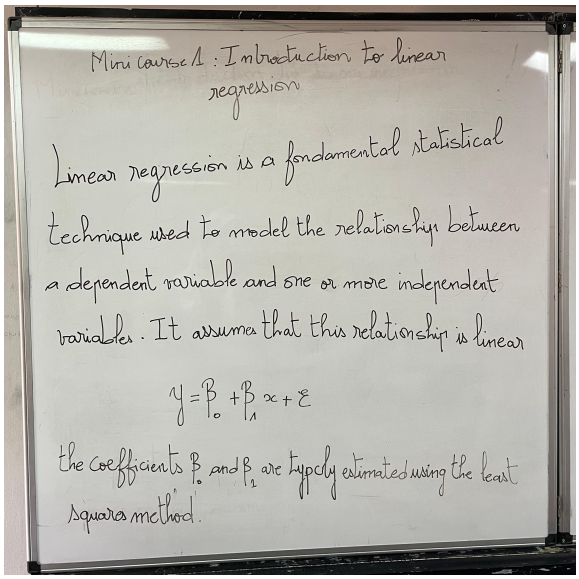


Fig. 2. Sample whiteboard image from the custom dataset, featuring handwritten text and mathematical equations under realistic classroom conditions.

C. Model

1) Traditional OCR systems: Traditional Optical Character Recognition (OCR) systems form the backbone of automated text extraction technology, employing classical computer vision techniques and machine learning algorithms to convert image-based text into machine-readable format. In this study, three prominent OCR engines were utilized to establish baseline performance for whiteboard text extraction.

a) PaddleOCR: PaddleOCR, developed by Baidu's PaddlePaddle team, integrates deep learning architectures with traditional text detection pipelines. The system employs a two-stage approach comprising text detection and recognition modules, utilizing convolutional neural networks optimized for multilingual character recognition. Its lightweight architecture offers computational efficiency while maintaining robust preprocessing capabilities for handling varied text orientations and lighting conditions [10].

b) Tesseract OCR: Tesseract OCR represents the industry standard in open-source OCR technology, originally developed by Hewlett-Packard and maintained by Google. The system combines traditional pattern recognition algorithms with LSTM neural networks, featuring a sophisticated pipeline that includes layout analysis, character segmentation, and statistical language modeling. Its modular architecture allows extensive customization through configuration parameters and supports multiple languages [11], [12].

c) EasyOCR: EasyOCR provides a streamlined deep learning implementation built on PyTorch, incorporating CRAFT (Character Region Awareness for Text detection) for text detection and CRNN (Convolutional Recurrent Neural Network) for character recognition. The system emphasizes out-of-the-box functionality with minimal configuration requirements while demonstrating particular strength in handling low-quality images and irregular text layouts through its modern neural network architectures [21].

2) LLM-Based API extraction: In parallel with traditional OCR techniques, this study investigates the effectiveness of modern Large Language Models (LLMs) for whiteboard content extraction. While OCR engines are designed to recognize textual patterns from images, they often fall short in comprehending structure, context, or ambiguous handwriting—especially in real-world conditions involving varied layouts, mathematical symbols, and inconsistent lighting. With the recent integration of vision capabilities into LLMs, Cognitive services and AI-based visual understanding platforms have further accelerated the integration of intelligent document analysis into real-world applications [26]. These models have become capable of interpreting not only plain text but also complex visual inputs such as scanned documents, handwritten notes, and whiteboard captures [21], [23].

Given their multi-modal capabilities, we explored whether LLMs could offer a superior alternative for digitizing educational content written on whiteboards, including text, equations, bullet points, and tables. Our objective was to benchmark the performance of these models using the same dataset developed for evaluating OCR tools, and assess whether LLMs could reliably serve as the core intelligence behind Clarix's text extraction pipeline [29].

For this evaluation, five state-of-the-art LLMs with multimodal capabilities were selected: GPT-4o (OpenAI), Claude 3 (Anthropic), Gemini 1.5 Pro (Google), DeepSeek-Vision (DeepSeek AI), and Grok (xAI). These models were chosen based on their publicly available visual understanding features, popularity, and architectural diversity. Rather than using the API endpoints of each model—which often involve complex setup procedures, account approval, quota limitations, or billing constraints—we opted to use their official user interfaces (web platforms) to ensure a consistent, hassle-free evaluation environment [22].

Each model was accessed through its respective interface (GPT-4o, Claude.ai, Gemini, DeepSeek chat, and X/Twitter AI chat), and tested under identical conditions using a standardized set of whiteboard images. This approach allowed us to focus purely on evaluating extraction performance without being affected by inconsistencies in API behavior or permission issues. It also reflects a real-world use case for educators or developers who might prefer visual input through UI platforms rather than programmatic pipelines. Despite not using APIs during this benchmark phase, our goal is to eventually integrate the most effective model into the Clarix system through API deployment, once a suitable model has been identified [28].

Although the use of public web interfaces simplified the evaluation process and ensured accessibility, this approach introduces certain reproducibility limitations. Consumer-facing interfaces may involve hidden system-level optimizations, dynamic model updates, session-state dependencies, and undocumented preprocessing operations that cannot be fully controlled experimentally. Consequently, slight variations in extraction behavior may occur across sessions or future platform versions. Future work will, therefore, focus on standardized API-based evaluation pipelines to improve experimental reproducibility and benchmarking consistency.

D. Evaluation

To evaluate the performance of Large Language Models (LLMs) and traditional OCR models in interpreting whiteboard content, we conducted a comparative analysis using five distinct LLM models: GPT-4o, Claude, Gemini, Deepseek, and Grok, alongside three traditional Optical Character Recognition (OCR) models: PaddleOCR, EasyOCR, and TesseractOCR. Each model was tested on a consistent dataset of five whiteboard images, covering various educational topics such as mathematics, physics, and economics.

The LLM models received identical prompts and were evaluated based on their ability to accurately reconstruct the textual content present in each image, leveraging their multimodal capabilities to understand both visual and textual elements within the whiteboard context. In contrast, the traditional OCR models relied purely on character recognition algorithms to extract text from the images.

For evaluation precision, recall, and F1 score were used as the primary metrics. These scores were computed by comparing the tokenized output of each model against manually defined ground truth references for each image. Precision measures the proportion of correctly identified tokens among all tokens predicted by the model, recall measures the proportion of correctly identified tokens among all actual tokens in the

ground truth, and F1 score represents the harmonic mean of precision and recall, providing a balanced assessment of model performance.

IV. Results

The comparative analysis revealed significant performance differences between Large Language Models (LLMs) and traditional OCR approaches in whiteboard content interpretation. Table I presents the average precision, recall, and F1 scores across all five whiteboard images for each model.

TABLE I. Average Precision, Recall, and F1 Score for each LLM and OCR Model.

Model	Avg Precision	Avg Recall	Avg F1 Score
Gemini	0.8188	0.8784	0.8466
GPT-4o	0.7954	0.8534	0.8162
Deepseek	0.7422	0.8625	0.7965
Grok	0.7372	0.7987	0.7645
Claude	0.7219	0.8034	0.7550
PaddleOCR	0.3523	0.3223	0.3333
EasyOCR	0.1101	0.1309	0.1158
TesseractOCR	0.0000	0.0000	0.0000

Fig. 3 presents a visual comparison of the F1 scores achieved by the evaluated Large Language Models (LLMs) and traditional OCR systems on the whiteboard text extraction benchmark.

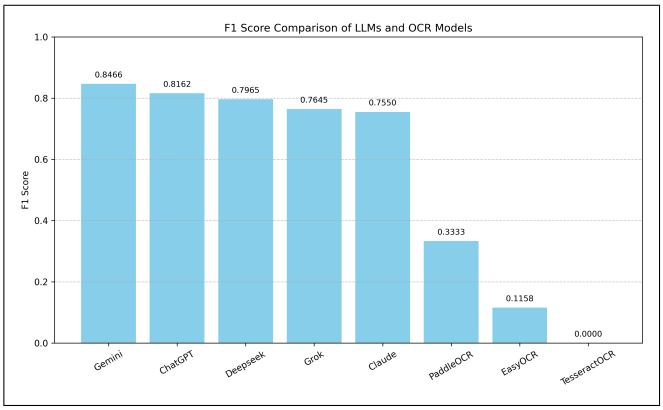


Fig. 3. F1 score comparison of various Large Language Models (LLMs) and OCRs on the text extraction benchmark.

Among the Large Language Models (LLMs), Gemini achieved the highest F1 score of 0.8466 with precision of 0.8188 and recall of 0.8784. GPT-4o ranked second with an F1 score of 0.8162, precision of 0.7954, and recall of 0.8534. Deepseek recorded the highest recall score of 0.8625 but achieved an F1 score of 0.7965 with precision of 0.7422. Grok obtained an F1 score of 0.7645 with precision of 0.7372 and recall of 0.7987. Claude achieved an F1 score of 0.7550 with precision of 0.7219 and recall of 0.8034.

Among the traditional OCR models, PaddleOCR achieved the highest performance with an F1 score of 0.3333, precision of 0.3523, and recall of 0.3223. EasyOCR recorded an F1 score of 0.1158 with precision of 0.1101 and recall of 0.1309. TesseractOCR recorded zero scores across all metrics.

Fig. 4, Fig. 5 and Fig. 6 display the text region detection capabilities of the three traditional OCR models. TesseractOCR detected 0 text regions across the whiteboard images.

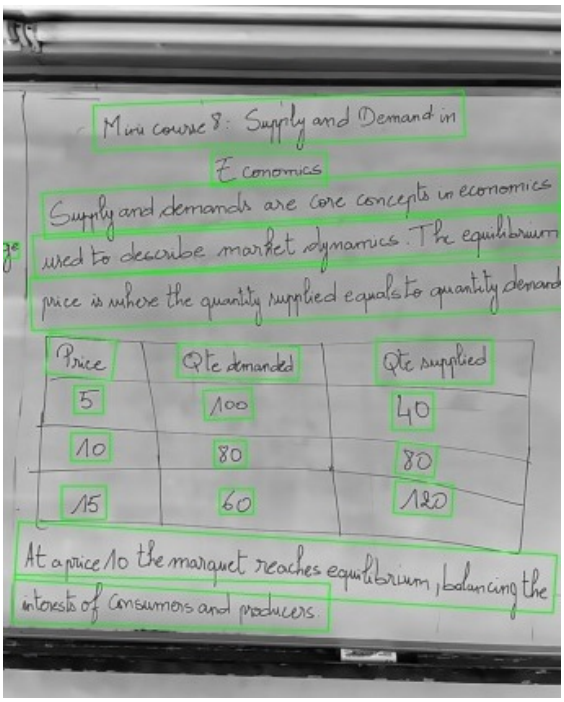


Fig. 4. PaddleOCR Optimized text region detection: 20 regions detected with blue bounding boxes covering complete phrases and text blocks.

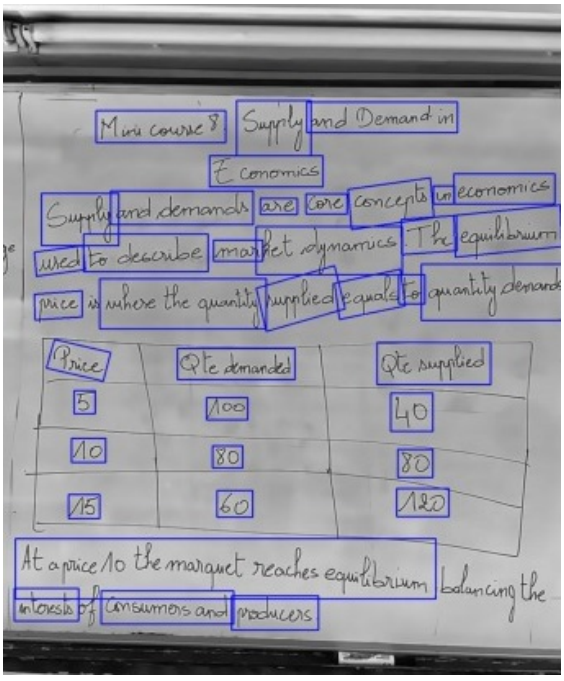


Fig. 5. EasyOCR text region detection: 39 regions detected with blue bounding boxes outlining individual words and text elements.

EasyOCR identified 39 text regions with blue bounding boxes outlining individual words and text elements. PaddleOCR Optimized detected 20 text regions with blue bounding boxes covering complete phrases and text blocks.

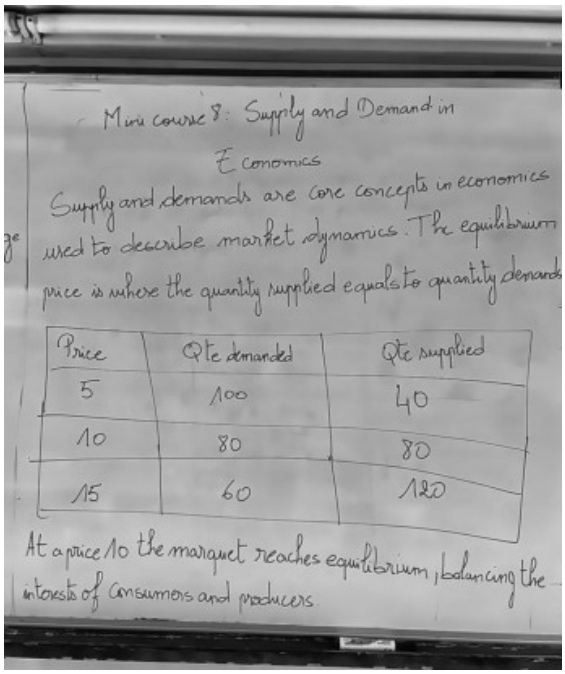


Fig. 6. TesseractOCR text region detection: 0 regions detected.

V. Discussion

The experimental results reveal substantial performance disparities between multimodal Large Language Models (LLMs) and traditional OCR systems when applied to whiteboard content extraction. The superior performance of LLMs across all evaluation metrics demonstrates the advantages of multimodal understanding capabilities over conventional character recognition approaches [14], [15].

Among the LLM models, Gemini's leading performance (F1 = 0.8466) can be attributed to its balanced precision and recall scores, indicating robust text extraction with minimal false positives and negatives. The relatively close performance clustering of all LLM models (F1 scores ranging from 0.7550 to 0.8466) suggests that current multimodal LLMs have reached a comparable level of maturity in visual text understanding tasks [14]. Notably, Deepseek achieved the highest recall score (0.8625), demonstrating superior capability in identifying existing text content, though this came at the cost of precision.

The stark performance gap between LLMs and traditional OCR models highlights fundamental limitations in conventional text recognition approaches when handling complex visual scenarios. PaddleOCR's relatively superior performance among OCR models (F1 = 0.3333) correlates with its text region detection strategy, which identifies 20 coherent text regions corresponding to complete phrases and sentences. This approach enables more contextual text extraction compared to word-level detection [18].

Conversely, EasyOCR's detection of 39 text regions reflects a granular, word-by-word approach that, while identifying more individual text elements, fails to maintain contextual relationships between words within phrases or sentences. This fragmented detection strategy likely contributes to its lower overall performance (F1 = 0.1158) despite detecting nearly twice as many regions as PaddleOCR. The word-level segmentation approach appears less effective for whiteboard scenarios where maintaining textual context and structure is crucial for accurate content reconstruction [14].

TesseractOCR's complete failure (0 detected regions, 0 scores across all metrics) indicates fundamental incompatibility with handwritten whiteboard content. This suggests that TesseractOCR's optimization for printed text renders it unsuitable for educational whiteboard scenarios containing handwritten mathematical expressions, diagrams, and informal notation styles [18].

The region detection analysis reveals that mere quantity of detected regions does not correlate with extraction quality. EasyOCR's higher region count paradoxically led to inferior performance compared to PaddleOCR's more selective but contextually coherent region identification. This finding underscores the importance of semantic understanding in text extraction tasks, where maintaining phrase-level and sentence-level coherence proves more valuable than exhaustive word-level detection [14].

The multimodal LLMs' consistent outperformance suggests that their training on diverse visual-textual datasets enables better handling of whiteboard-specific challenges, including handwritten text, mathematical notation, informal layouts, and mixed content types. Unlike traditional OCR systems that rely solely on pattern matching and character recognition algorithms, LLMs can leverage contextual understanding to infer missing or unclear characters and maintain semantic coherence across extracted text [15], [17].

These findings have practical implications for educational technology applications, where accurate whiteboard content extraction is essential for automated note-taking, content digitization, and accessibility enhancement. The results suggest that current multimodal LLMs provide viable solutions for such applications, while traditional OCR approaches remain inadequate for complex whiteboard scenarios [16].

The experimental observations also revealed several recurring OCR failure modes. Traditional OCR systems frequently struggled with handwritten mathematical symbols, overlapping characters, irregular spacing, marker reflections, and non-linear text arrangements commonly found on whiteboards. In particular, equations containing fractions, superscripts, subscripts, and mixed alphanumeric notation often produced fragmented or incomplete extraction results. OCR systems also exhibited sensitivity to low-contrast regions and partially erased text, which negatively impacted segmentation accuracy and contextual reconstruction.

Compared with previously reported whiteboard automation systems, Clarix introduces a more comprehensive integration of embedded automation and intelligent digitization. Earlier systems primarily focused either on automated surface cleaning mechanisms or on standalone OCR-based text extraction without semantic understanding capabilities. In contrast, Clarix combines real-time image acquisition, automated erasure, and multimodal Large Language Model-based content interpretation within a unified low-cost embedded platform. This integration enables not only physical automation but also contextual reconstruction and digital archiving of handwritten

educational content, thereby extending the functionality of traditional whiteboard automation approaches.

This study presents several limitations that should be considered when interpreting the results. First, the evaluation dataset remains relatively small and limited to five whiteboard images captured under controlled conditions. Although the selected samples cover different educational domains, a substantially larger dataset would be necessary for statistically robust benchmarking. Second, the experiments primarily focused on English handwritten educational content, which may limit generalization to multilingual environments and highly diverse handwriting styles. Third, the evaluated multimodal Large Language Models (LLMs) were tested through public user interfaces rather than direct API integration, potentially introducing variations in response behavior and latency. In addition, the current study does not extensively evaluate robustness against highly degraded writing conditions, extreme lighting variations, or large-scale classroom deployments. Future work will focus on expanding the dataset, improving real-time processing capabilities, and evaluating the system across broader educational scenarios and handwriting conditions.

Future research directions should explore the specific factors contributing to LLM superiority, investigate hybrid approaches combining OCR region detection with LLM contextual understanding, and evaluate performance across diverse whiteboard content types and writing styles [14].

VI. Conclusion

Clarix demonstrates an effective, low-cost solution for automating the digitization and cleaning of whiteboard content using embedded systems and AI technologies. By integrating a Raspberry Pi, OCR, and a servo-driven cleaning mechanism, the system bridges the gap between analog and digital workflows. The evaluation of five advanced multimodal LLMs on diverse whiteboard content provides valuable insights into their performance in real-world educational scenarios. Gemini achieved the highest overall performance (F1=0.8466), demonstrating the best balance between precision (0.8188) and recall (0.8784), while GPT-4o ranked second (F1=0.8162) and DeepSeek recorded the highest recall score (0.8625). All LLMs substantially outperformed traditional OCR engines, with PaddleOCR being the only conventional system to achieve meaningful extraction results (F1=0.3333), and TesseractOCR failing entirely on handwritten whiteboard content. Clarix not only streamlines classroom and meeting room documentation but also lays the groundwork for scalable deployments in educational and collaborative settings. Future work will explore real-time transcription, multi-language support, and cloud integration to further enhance the system's capabilities. Based on the obtained results, Gemini appears to be the most suitable candidate for future API integration.

References

[1] Uduak, Idoghor, and R. Kasumu, "The use of interactive whiteboards for teaching and learning in tertiary institutions," *International Journal of Trendy Research in Engineering and Technology*, vol. 6, pp. 28–33, Jan. 2022, doi: 10.54473/IJTRET.2022.6606.

[2] O. Onunka, T. Onunka, A. Fawole, I. Adeleke, and C. Daraojimba, "Library and information services in the digital age: Opportunities and challenges," *Acta Informatica Malaysia*, vol. 7, pp. 113–121, Sep. 2023, doi: 10.26480/aim.02.2023.113.121.

[3] B. U. Nwozor and O. F. Obichie, "Real-time artificial intelligence system for collaborative document editing and workload optimisation," *FUPRE Journal of Scientific & Industrial Research*, vol. 9, no. 3, pp. 282–299, Jul. 2025.

[4] S. Mhlongo, K. Mbatha, B. Ramatsetse, and R. Dlamini, "Challenges, opportunities, and prospects of adopting and using smart digital technologies in learning environments: An iterative review," *Heliyon*, vol. 9, no. 6, 2023, doi: 10.1016/j.heliyon.2023.e16348.

[5] S. H. Qasim, *Beyond the Classroom: Emerging Technologies to Enhance Learning*. Book Bazooka Publication, 2024.

[6] Y. Jayaram and J. Bhat, "Intelligent forms automation for higher education: Streamlining student onboarding and administrative workflows," *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 3, no. 4, pp. 100–111, Dec. 2022, doi: 10.63282/3050-9246.IJETCSIT-V3I4P110.

[7] D. Ali *et al.*, "Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections," *Journal of Documentation*, vol. 80, no. 5, pp. 1031–1056, Feb. 2023, doi: 10.1108/JD-01-2022-0029.

[8] E. J. Chukwuemeka, "Smart education: Opportunities, challenges and future of traditional education," *International Journal of Smart Technology and Learning*, vol. 4, no. 3, pp. 191–202, 2025, doi: 10.1504/IJSMARTTL.2025.146286.

[9] M. Alenezi, S. Wardat, and M. Akour, "The need of integrating digital education in higher education: Challenges and opportunities," *Sustainability*, vol. 15, no. 6, p. 4782, 2023, doi: 10.3390/su15064782.

[10] O. Sarkar, S. Sinha, A. K. Jena, A. K. Parida, N. Parida, and R. K. Parida, "Automatic Number Plate Character Recognition using PaddleOCR," in *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*, pp. 1–7, Jun. 2024, doi: 10.1109/ICICET59348.2024.10616305.

[11] T. Malathi, D. Selvamuthukumaran, C. S. Diwaan Chandar, V. Niranjan, and A. K. Swashthika, "An Experimental Performance Analysis on Robotics Process Automation (RPA) With Open Source OCR Engines: Microsoft OCR And Google Tesseract OCR," *IOP Conference Series: Materials Science and Engineering*, vol. 1059, no. 1, p. 012004, Feb. 2021, doi: 10.1088/1757-899X/1059/1/012004.

[12] N. Anakpluek, W. Pasanta, L. Chantharasukha, P. Chokratansombat, P. Kanjanakaew, and T. Siriborvornratanakul, "Improved Tesseract optical character recognition performance on Thai document datasets," *Big Data Research*, vol. 39, p. 100508, Feb. 2025, doi: 10.1016/j.bdr.2025.100508.

[13] P. Jain, K. Taneja, and H. Taneja, "Which OCR toolset is good and why: A comparative study," *Kuwait Journal of Science*, vol. 48, no. 2, Apr. 2021, doi: 10.48129/kjs.v48i2.9589.

[14] F. Seitl, T. Kovářík *et al.*, "Assessing the quality of information extraction," *arXiv preprint arXiv:2404.04068*, 2024.

[15] A. Grothey *et al.*, "Comprehensive testing of Large Language Models (LLMs) for extraction of structured pathology information," *Nature Communications*, 2025.

[16] J. Wang and L. Jing, "A Virtual Reality Whiteboard System for Remote Collaboration Using Natural Handwriting," *Electronics*, vol. 11, no. 24, p. 4152, Jan. 2022, doi: 10.3390/electronics11244152.

[17] M. Schilling-Wilhelmi *et al.*, "From text to insight: Large Language Models (LLMs) for chemical data extraction," *Chemical Society Reviews*, vol. 54, no. 3, pp. 1125–1150, 2025, doi: 10.1039/D4CS00913D.

[18] A. Al Zubaer, M. Granitzer, and J. Mitrović, "Performance analysis of Large Language Models (LLMs) in the domain of legal argument mining," *Frontiers in Artificial Intelligence*, vol. 6, 2023, doi: 10.3389/frai.2023.1278796.

[19] S. Muharom *et al.*, "Detection and erasing scribble blackboard system based on Hough-transform method using camera," *Journal of Physics: Conference Series*, vol. 2117, no. 1, p. 012010, 2021, doi: 10.1088/1742-6596/2117/1/012010.

[20] M. D. Alahmadi and M. Alshangiti, "Optimizing OCR performance for programming videos: The role of image super-resolution and Large Language Models (LLMs)," *Mathematics*, vol. 12, no. 7, p. 1036, 2024, doi: 10.3390/math12071036.

[21] O. H. Abdellaif, A. N. Hassan, and A. Hamdi, "ERPA: Efficient RPA Model Integrating OCR and LLMs for Intelligent Document Processing," in *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 295–300, Nov. 2024, doi: 10.1109/MIUCC62295.2024.10783599.

[22] K. Khanchandani *et al.*, "Automated invoice data extraction: Using LLM and OCR," *arXiv preprint arXiv:2511.05547*, 2026.

[23] S. Kim *et al.*, "Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records," *arXiv preprint arXiv:2501.11623*, 2025.

[24] N. Y. WinWar and Z. Myo Tun, "IR controlled automatic whiteboard cleaner using Arduino," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 8, no. 6, Jun. 2019.

[25] F. Naiemi, V. Ghods, and H. Khalesi, "Scene text detection and recognition: A survey," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 20255–20290, 2022, doi: 10.1007/s11042-022-12693-7.

[26] A. Moniz, M. Gordon, I. Bergum, M. Chang, and G. Grant, "Introducing cognitive services," in *Beginning Azure Cognitive Services: Data-Driven Decision Making Through Artificial Intelligence*. Berkeley, CA: Apress, 2021, pp. 1–17, doi: 10.1007/978-1-4842-7176-6_1.

[27] G. Nunes, V. Rolla, D. Pereira, V. Alves, A. Carreiro, and M. Baptista, "Benchmarking table extraction: Multimodal LLMs vs traditional OCR," in *Proceedings of the 1st Joint Workshop on Large Language Models (LLMs) and Structure Modeling (XLLM 2025)*, Vienna, Austria: Association for Computational Linguistics, Aug. 2025, pp. 8–15, doi: 10.18653/v1/2025.xllm-1.2.

[28] C. J. Pieterse and P. Venter, "An OCR-LLM framework towards tabular data extraction," in *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Cham: Springer, 2026, pp. 176–188, doi: 10.1007/978-3-032-08384-5_15.

[29] T. Doan and R. Koynov, *Opportunities and challenges of LLMs as post-OCR correctors*, vol. 45, Oct. 2025, doi: 10.15439/2025F4697.

Adaptive Temporal Windowing for Streaming Outlier Detection Under Dynamic Arrival Rates

Hend Maher[1], Mohamed Khafagy[2], Heba Nagaty[3]
Faculty of Computers and Artificial Intelligence, Department of Information Systems, Fayoum University, Egypt[1,2,3]
Faculty of Computer Science and Information Systems, Department of Information Systems,
October 6 University, Egypt[1]

***Abstract*—Streaming outlier detection requires adaptive mechanisms capable of handling continuously evolving data streams under dynamic arrival rates. Existing count-based approaches fail under bursty and irregular stream arrival patterns commonly observed in real-world systems, since they trigger model updates after a fixed number of instances without considering temporal dynamics. In this study, we propose an adaptive time-driven windowing framework for streaming outlier detection that decouples model updates from instance count and instead leverages elapsed time as the primary control mechanism. The proposed approach is based on the Density Incremental Local Outlier Factor (DILOF) and introduces a time-aware update strategy aligned with real-world streaming behavior. Extensive experiments on benchmark datasets demonstrate that the proposed method achieves robust and stable detection performance, with AUC values ranging up to 0.96. The results further show that time-based windowing provides a consistent trade-off between detection accuracy and computational efficiency, while offering a temporally grounded update mechanism for streams with variable arrival behavior. In addition, we analyze hybrid count-time strategies and demonstrate their limitations due to dominance effects. Repeated runs further indicate the robustness and consistency of the proposed framework. The findings highlight that temporal awareness is a critical factor in stream outlier detection and should be explicitly incorporated into windowing mechanisms, particularly in resource-constrained environments such as fog and edge computing.**

Keywords*—Streaming outlier detection; time-based windowing; Density Incremental Local Outlier Factor; data streams; anomaly detection; fog computing

I. Introduction

Outlier detection in data streams has become increasingly important in many real-world applications, such as network monitoring, financial fraud analysis, industrial process control, and sensor-based systems. In these scenarios, data is generated continuously and often at high speed, making it impractical to store and reprocess the entire stream [1], [2]. As a result, outlier detection methods that stream must operate incrementally while maintaining both efficiency and detection accuracy.

As illustrated in Fig. 1, outliers are typically characterized by their presence in low-density regions compared to normal data points, which tend to form dense clusters [3]. This observation forms the basis of density-based methods such as LOF, which evaluate the relative density of each data point with respect to its neighbors [4], [5]. Furthermore, the effectiveness of density-based outlier detection is closely related to how the local neighborhood is constructed and updated over time [8], [9]. In streaming environments, where data

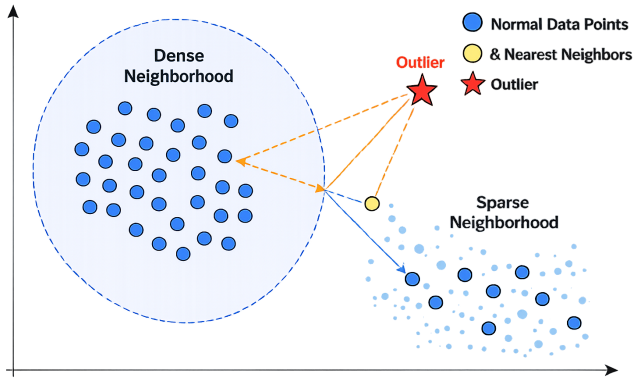


Fig. 1. Conceptual illustration of density-based outlier detection, where normal data points form dense regions while outliers reside in sparse regions with significantly lower local density.

arrives continuously, maintaining an accurate representation of these density regions depends on the chosen windowing mechanism [6]. This motivates the investigation of time-based windowing as a more adaptive approach for capturing the evolving structure of the data stream [10].

A central design issue in outlier detection of the stream is how to manage the active data window used for model updating. Most existing methods rely on count-based windowing, where the system updates its internal representation after processing a fixed number of instances [7], [12]. Although this strategy is widely used, it assumes that the number of arriving instances alone is sufficient to regulate the update process. However, in practice, data streams may exhibit irregular arrival rates, bursts of activity, or periods of relative sparsity. Under such conditions, a purely count-based mechanism may not reflect the actual temporal behavior of the stream.

These observations motivate the study of time-based windowing. Instead of triggering summarization after a fixed number of instances, time-based windowing uses elapsed time as the controlling factor. This allows the update process to become more sensitive to the temporal dynamics of the stream and may improve the trade-off between model freshness and computational cost.

In this study, we investigate the role of time-based windowing in streaming outlier detection using Density Incremental Local Outlier Factor (DILOF) as a baseline framework. We

implement a time-driven variant and evaluate its behavior on benchmark datasets. Our experiments show that the time-based strategy achieves strong and stable performance, while the optimal time window varies across datasets, suggesting that time-window selection should be data-dependent rather than fixed universally.

In addition to the time-based variant, we explore several hybrid strategies that combine count-based and time-based triggers. These include rule-based, score-based, and adaptive formulations. However, our empirical results indicate that simple hybridization mechanisms often become dominated by one trigger, preventing a genuinely balanced behavior. This finding highlights that combining count and time in streaming settings requires more principled formulations than straightforward trigger fusion.

The problem is also relevant in fog computing and edge environments, where data must often be processed locally with limited resources and low latency requirements [14], [15]. In such settings, an effective time-aware summarization policy can be valuable for maintaining timely anomaly detection without excessive computational overhead [17].

The main contributions of this study are summarized as follows:

- We introduce an adaptive time-driven windowing framework that redefines update triggering in streaming outlier detection by explicitly incorporating temporal dynamics.
- We develop a time-based variant of the Density Incremental Local Outlier Factor (DILOF) and evaluate its effectiveness on benchmark datasets with varying arrival behavior.
- We demonstrate that time-based windowing achieves robust and stable detection performance while providing a better balance between responsiveness and computational efficiency.
- We provide both theoretical and empirical analysis of time-based windowing and explain its advantages over conventional count-based mechanisms under dynamic arrival rates.
- We analyze hybrid count-time strategies and reveal their practical limitations due to trigger dominance effects.
- We discuss the applicability of the proposed framework in real-world streaming environments, including IoT, industrial monitoring, fog computing, and edge systems.

Despite extensive research in outlier detection, most existing methods still rely on instance-driven update mechanisms that fail to capture temporal dynamics in real-world data streams. This limitation leads to suboptimal update behavior under irregular arrival patterns, including bursty and sparse regimes. Unlike conventional approaches that merely adjust fixed count thresholds, the proposed framework redefines update triggering through temporal stream behavior, enabling adaptive model refresh under fluctuating data rates. By decoupling summarization from instance count, the proposed time-driven strategy provides a more realistic and practically meaningful mechanism for maintaining detection quality in streaming environments.

II. Related Work

Streaming outlier detection has been extensively studied in the context of real-time data processing, where models must operate incrementally under limited memory and computational constraints [16]. Early approaches focused on distance-based methods [8], [9], [11], which identify anomalies based on their distance from neighboring instances. However, these methods often suffer from scalability limitations in high-volume data streams.

Density-based approaches, particularly the Local Outlier Factor (LOF) [18], have gained significant attention due to their ability to capture local data characteristics. Several extensions have been proposed to adapt LOF to streaming environments, including incremental and memory-efficient variants, such as DILOF [13] and EILOF [23]. These methods improve efficiency by updating local density estimates incrementally while maintaining a sliding window of recent instances.

A key design component in streaming algorithms is the windowing mechanism used to control model updates. Most existing approaches rely on count-based windowing, where summarization is triggered after processing a fixed number of instances. While effective in maintaining bounded memory, this strategy assumes a uniform data arrival rate, which is often unrealistic in real-world scenarios characterized by bursty or irregular data flows.

Recent studies have explored adaptive and hybrid windowing strategies to address this limitation. However, many of these approaches still rely on instance-based triggers or simple combinations of count and time, which may not fully capture the temporal dynamics of streaming data [18],[20], [13].

In addition to LOF-based streaming methods, ensemble-based anomaly detection methods such as Isolation Forest have been widely adopted due to their efficiency and scalability. However, such methods are not specifically designed to incorporate temporal windowing mechanisms in streaming environments, and therefore do not explicitly address the challenges associated with dynamic data arrival patterns.

Despite these advancements, most existing approaches continue to rely on instance-driven update mechanisms, limiting their ability to adapt to temporal variations in real-world data streams. This highlights a critical gap in current research.

Table I summarizes the positioning of the proposed TD-ILOF framework relative to representative outlier detection methods of streaming. Existing approaches have made important progress in improving memory efficiency, incremental processing, and adaptability to evolving streams. However, their update mechanisms are generally driven by instance arrivals, adaptive context rules, or model-specific online updates, rather than by elapsed time as an explicit control variable for summarization. In contrast, TD-ILOF focuses on the role of time-driven windowing in density-based streaming outlier detection and systematically examines its behavior against conventional count-based and hybrid alternatives.

TABLE I. COMPARATIVE POSITIONING OF REPRESENTATIVE STREAMING OUTLIER DETECTION METHODS

Method	Year	Main Idea	Update / Windowing Mechanism	Temporal Adaptation	Key Limitation
DILOF [13]	2018	Density-based local outlier detection with memory-efficient summarization	Incremental LOF updates with instance-driven summarization	Does not explicitly use elapsed time as the primary update trigger	Assumes updates are governed mainly by processed instance counts
ASOD [6]	2024	Adaptive online stream outlier detection using context-aware modeling	Online incremental updates with adaptive context control	Addresses stream evolution and contextual adaptation	Does not focus on time-driven windowing for density-based summarization
EILOF [23]	2025	Efficient incremental LOF approximation for data streams	Updates LOF-related quantities primarily for newly arriving points	Improves online processing efficiency	Does not study time-based summarization as a control mechanism
EiForestASD [24]	2025	Tree-based streaming anomaly detection derived from Isolation Forest	Stream-oriented ensemble update strategy	Designed for efficient online anomaly detection	Not density-based and does not analyze temporal windowing in LOF-style summarization
TD-ILOF (Ours)	**2026**	**Density-based streaming outlier detection with time-driven summarization**	**Time-based, count-based, and hybrid update strategies**	**Explicitly models elapsed time as a primary update trigger**	**Targets the gap of temporal update control in density-based stream detection**

To address this limitation, this work focuses on time-based windowing as a primary mechanism for controlling model updates. By decoupling the update process from instance count, the proposed approach adapts naturally to variations in data arrival rates. To the best of our knowledge, a systematic empirical and theoretical evaluation of time-driven windowing in the context of density-based streaming outlier detection remains limited.

Unlike event-time processing frameworks, which organize stream computation according to timestamps attached to arriving records, the present work focuses on controlling the summarization schedule of a density-based outlier detector through elapsed processing time. Similarly, existing adaptive windowing studies usually aim to revise window boundaries or update policies at the stream-system level, whereas TD-DILOF targets the internal refresh mechanism of DILOF itself. Therefore, the contribution of this work is not a generic time-windowing replacement, but a detector-level temporal control strategy for density-preserving streaming summarization.

This study aims to bridge this gap by analyzing the effectiveness of time-based windowing and comparing it with conventional count-based and hybrid strategies.

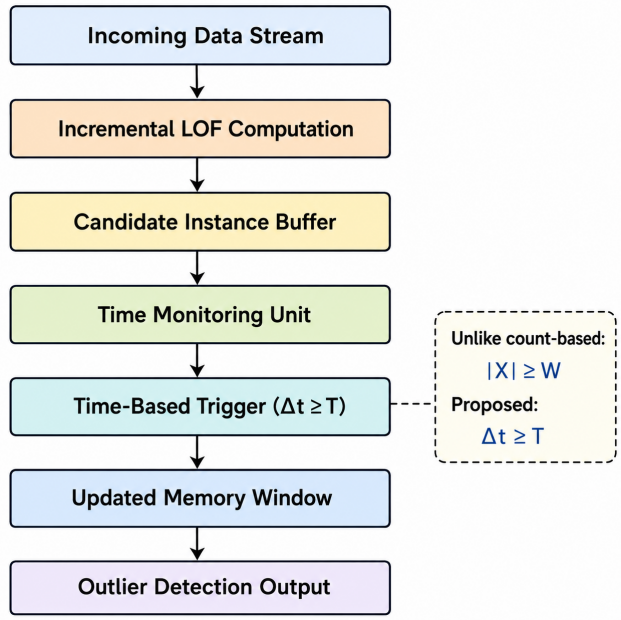


Fig. 2. Framework of the proposed time-based streaming outlier detection approach. Incoming data is processed incrementally and summarized based on elapsed time.

III. METHODOLOGY

In this section, we describe the baseline streaming outlier detection method and the proposed time-based windowing strategy.

A. Background on LOF-Based Methods

Streaming outlier detection aims to assign an anomaly score to each incoming data point in a continuous data stream $P = \{p_1, p_2, \dots\}$ [19], [22]. This process must operate under strict constraints, including limited memory capacity, real-time processing requirements, and evolving data distributions [26], [25].

A widely adopted technique for outlier detection is the Local Outlier Factor (LOF) algorithm, which evaluates the degree of abnormality of a data point based on its local density relative to its neighbors.

Let k denote the number of nearest neighbors, $dist(x, y)$ the Euclidean distance between two points, and $N_k(x)$ the set of k nearest neighbors of x. The computation of LOF is defined as follows.

Definition 1. The reachability distance is given by:

$$reach_dist_k(x, y) = \max\{dist(x, y), dist_k(y)\} \quad (1)$$

Definition 2. The local reachability density is:

$$lrd_k(x) = \left(\frac{1}{k} \sum_{y \in N_k(x)} reach_dist_k(x, y) \right)^{-1} \quad (2)$$

Definition 3. The LOF score is:

$$LOF_k(x) = \frac{1}{k} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)} \quad (3)$$

A data point is considered an outlier if its LOF score exceeds a predefined threshold.

Several extensions have been proposed to adapt LOF to streaming environments. The incremental LOF (iLOF) processes incoming data sequentially but requires storing all instances, which limits scalability. MiLOF addresses this limitation through clustering-based summarization but may lose important density information. More recently, DILOF improves upon these approaches by preserving density structures during summarization, making it a suitable baseline for streaming outlier detection. As shown in Fig. 2, the proposed approach replaces the traditional count-based trigger with a time-based mechanism.

B. Baseline: Density Incremental Local Outlier Factor

Density Incremental Local Outlier Factor (DILOF) is an incremental extension of the Local Outlier Factor (LOF) algorithm designed for streaming data. Instead of processing the entire dataset at once, DILOF updates the outlier scores incrementally as new instances arrive.

The method maintains a sliding window of recent data points and computes the local density of each instance based on its k-nearest neighbors. The LOF score is then used to determine whether a point is an outlier relative to its local neighborhood.

In the standard configuration, DILOF relies on a count-based windowing mechanism. A fixed window size W is used, and once the number of instances in the window reaches W, a summarization process is triggered. During summarization, a subset of representative instances is retained while less informative instances are removed to control memory usage and computational cost.

C. Count-Based Windowing

In the count-based approach, summarization is triggered when the number of processed instances reaches a predefined threshold W. Formally, let N denote the number of instances currently stored in the window. The summarization condition is defined as:

$$N \geq W \quad (4)$$

This mechanism ensures a bounded memory footprint but does not consider the temporal properties of the data stream. As a result, the update frequency depends solely on the number of arriving instances, which may lead to suboptimal behavior when the data arrival rate varies over time.

D. Proposed Time-Driven Density Incremental Local Outlier Factor

To address the limitations of count-based windowing, we propose a time-based mechanism in which summarization is triggered according to elapsed time rather than instance count.

Let T denote the predefined time window in milliseconds, and let Δt represent the elapsed time since the last summarization. The summarization condition is defined as:

$$\Delta t \geq T \quad (5)$$

In this approach, each incoming instance is processed incrementally, and a timer is used to track the elapsed time. Once the time threshold is reached, the summarization procedure is invoked, and the timer is reset. The distinction is therefore deeper than replacing the condition $N \geq W$ with $\Delta t \geq T$. In the proposed framework, the temporal trigger changes the rhythm of memory renewal and directly affects how many instances participate in each summarization cycle. This alters the interaction between stream arrival behavior, density preservation, and computational cost, which is not captured by a simple trigger substitution view. This mechanism allows the update frequency to adapt naturally to the temporal dynamics of the stream. In high-rate periods, more instances are processed within the same time interval, while in low-rate periods, fewer updates occur. This behavior provides a better balance between responsiveness and computational efficiency.

E. Hybrid Windowing Strategies

In addition to the time-based approach, we explored hybrid strategies that combine count-based and time-based triggers. These strategies include rule-based (logical OR), score-based combinations, and adaptive selection mechanisms.

However, our empirical evaluation shows that these naive hybrid formulations tend to be dominated by a single trigger, either count or time, depending on the configuration. As a result, they fail to provide a truly balanced integration of both mechanisms. This observation suggests that more principled approaches are required for effective hybrid windowing in streaming settings.

F. Theoretical Analysis of Time-Based Windowing

While time-based windowing provides an intuitive alternative to count-based mechanisms, it is important to analyze its behavior from a theoretical perspective in order to understand its impact on streaming outlier detection.

In count-based windowing, the update frequency is determined by the number of incoming instances. Assuming a fixed window size W, the expected update interval depends directly on the data arrival rate $r(t)$. When the arrival rate varies over time, the effective update interval becomes unstable, leading to inconsistent model refresh behavior.

In contrast, time-based windowing decouples the update mechanism from the instance count and instead relies on a fixed temporal interval T. Let N_T denote the number of instances observed within a time window T. Then:

$$N_T = \int_{t_0}^{t_0+T} r(t)\, dt \tag{6}$$

This formulation shows that the number of processed instances becomes a function of the arrival rate rather than a fixed constraint. As a result, the model adapts naturally to the dynamics of the data stream.

From a density estimation perspective, this behavior has important implications. During high-rate periods, the model receives more data within the same time window, leading to more accurate local density estimation. Conversely, during low-rate periods, fewer updates occur, reducing unnecessary computation while preserving temporal consistency.

G. Computational Complexity Analysis

The computational complexity of the proposed approach is influenced by both the LOF computation and the summarization process.

For each incoming instance, the LOF computation requires identifying the k-nearest neighbors, which typically has a complexity of $O(k \log n)$ depending on the data structure used.

In the count-based approach, summarization is triggered every W instances, resulting in a total of $O(N/W)$ summarization operations for a stream of size N.

In the time-based approach, the number of summarization operations depends on the total elapsed time rather than the number of instances. Let T_{total} denote the total processing time. Then the number of summarizations is approximately:

$$\frac{T_{total}}{T} \tag{7}$$

This leads to a more predictable and stable update frequency, particularly in environments with fluctuating data rates.

Moreover, as observed in the experimental results, increasing the time window reduces the number of summarization operations, which directly impacts execution time and computational cost.

H. Design Rationale and Trade-offs

The choice of time-based windowing introduces a fundamental trade-off between detection accuracy, responsiveness, and computational efficiency.

Smaller time windows result in more frequent updates, improving model freshness but increasing computational overhead. Larger time windows reduce the update frequency, improving efficiency but potentially delaying the detection of new anomalies.

This trade-off can be interpreted as a balance between temporal resolution and computational cost. Unlike count-based methods, which assume uniform data arrival, time-based windowing provides a mechanism that aligns more closely with real-world streaming conditions.

These characteristics make time-based windowing particularly suitable for deployment in resource-constrained environments such as fog and edge computing, where both latency and computational efficiency are critical.

1) Adaptive time window selection strategy: The experimental results indicate that the effectiveness of time-based windowing is influenced by the choice of the time window T, which appears to be dataset-dependent. While fixed time-window configurations provide stable performance, selecting an appropriate value of T remains a practical challenge, particularly in streaming environments with varying data arrival patterns.

To address this limitation, we outline a potential adaptive time window selection strategy that dynamically adjusts the window size based on the observed arrival rate of the data stream.

a) Motivation: In real-world streaming scenarios, the data arrival rate may fluctuate over time due to bursts, delays, or irregular generation processes. Under such conditions, a fixed time window may lead to inconsistent behavior. During high-rate periods, a fixed T may accumulate a large number of instances before summarization, increasing computational overhead. Conversely, during low-rate periods, the same window may contain too few instances, reducing the reliability of density estimation.

These observations suggest that adjusting the time window according to the temporal characteristics of the stream may improve both efficiency and stability.

b) Adaptive formulation: Let $r(t)$ denote the estimated data arrival rate at time t, defined as the number of instances observed per unit time. The adaptive time window $T(t)$ can be defined as:

$$T(t) = \alpha \cdot \frac{1}{r(t)} \tag{8}$$

where, α is a scaling factor controlling the balance between update frequency and computational cost.

Under this formulation, higher arrival rates lead to smaller time windows, resulting in more frequent updates. In contrast, lower arrival rates produce larger windows, reducing unnecessary computations. This behavior helps maintain a more consistent effective window content over time.

c) Integration with the framework: The adaptive mechanism can be incorporated into the proposed time-based framework by updating the time window prior to each summarization step:

```
if adaptive_mode:
    estimate r(t) over recent interval
    update T(t) = alpha / r(t)
```

```
if Delta_t >= T(t):
    perform summarization
    reset timer
```

In practice, the arrival rate $r(t)$ can be estimated using a sliding time interval or an exponential moving average to reduce sensitivity to short-term fluctuations.

d) Discussion: Although this adaptive strategy is not experimentally evaluated in the current study, it represents a natural extension of the proposed time-based windowing approach. It provides a principled direction for reducing sensitivity to dataset-specific configurations and may improve robustness in real-world streaming environments. Future work will focus on evaluating this strategy under real-time conditions and exploring more advanced adaptive formulations. Algorithm 1 presents the Time-Based Density Incremental Local Outlier Factor.

Algorithm 1 Time-Based Density Incremental Local Outlier Factor

Require: Infinite data stream $P = \{p_1, p_2, \dots, p_t, \dots\}$, time window T, LOF threshold θ, step size η, regularization constant λ, maximum number of iterations I
Ensure: Set of detected outliers O

$X \leftarrow \emptyset$ ▷ Data points in memory
$O \leftarrow \emptyset$ ▷ Detected outliers
startTime $\leftarrow$ *currentTime*()
for each $p_t \in P$ **do**
  $LOF_k(p_t) \leftarrow LOD(p_t, O, \theta)$
  if $LOF_k(p_t) > 0$ **then**
    $X \leftarrow X \cup \{p_t\}$
    $\Delta t \leftarrow$ *currentTime*() $-$ *startTime*
    if $\Delta t \geq T$ **then**
      $Z \leftarrow NDS(X, \eta, \lambda, I)$
      Remove the oldest $|X|/2$ data points from X
      $X \leftarrow X \cup Z$
      startTime $\leftarrow$ *currentTime*()
    end if
  end if
end for

IV. Experimental Setup

In this section, we describe the experimental configuration used to evaluate the proposed time-based windowing strategy.

A. Datasets

To assess the effectiveness of the proposed approach, we conducted experiments on two benchmark datasets commonly used in outlier detection research. These datasets differ in size, dimensionality, and anomaly construction, providing a suitable basis for evaluating the proposed method under different structural and computational conditions.

All datasets were processed in a simulated streaming environment, where instances were fed sequentially to emulate data arrival over time. No preprocessing was applied other than feature normalization, ensuring a fair and consistent evaluation across all methods.

TABLE II. Characteristics of the Benchmark Datasets used in the Experiments.

Characteristic	UCI Vowel	UCI Pendigit
Instances	1,456	3,498
Dimensions	12	16
Classes	11	10

Table II summarizes the main characteristics of the datasets used in the experiments.

The UCI Vowel and UCI Pendigit datasets were selected because they are commonly used benchmarks for evaluating density-based outlier detection methods. They also differ in size and dimensionality, allowing the proposed time-driven update policy to be examined under different data characteristics. In the Vowel dataset, one class was treated as the outlier class. In the Pendigit dataset, uniformly distributed noise was injected to simulate anomalous observations.

B. Evaluation Metric

The performance of the proposed method is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC). AUC is a widely used metric for outlier detection as it captures the trade-off between true positive rate (TPR) and false positive rate (FPR) across different threshold settings. Higher AUC values indicate better discrimination between normal instances and outliers.

C. Parameter Settings

For the time-based windowing strategy, different time windows ranging from 100 ms to 200 ms were evaluated. This range was selected to capture both frequent-update and moderate-update regimes in the simulated streaming environment, enabling analysis of the trade-off between responsiveness and computational overhead. The selected values provide a practical sensitivity analysis rather than assuming a universally optimal temporal configuration.

D. Implementation Details

All experiments were performed using an incremental streaming setup, where instances are processed sequentially in the order of arrival. The summarization process is triggered based on the selected windowing mechanism, either count-based or time-based.

For the time-based approach, a timer is used to measure the elapsed time between successive summarization operations. Once the predefined time threshold is reached, summarization is executed and the timer is reset. No artificial delays were introduced during the experiments, ensuring that the results reflect the natural processing time of the system. All experiments were conducted on a standard computing environment. Accordingly, the reported timing behavior should be interpreted as processing-time-triggered summarization within a controlled experimental environment, rather than as a full event-time implementation for distributed stream-processing systems.

E. Comparison Strategy

To evaluate the effectiveness of the proposed approach, we compare the time-driven windowing strategy with the traditional count-based mechanism and representative LOF-based streaming baselines. This comparison highlights the impact of incorporating temporal information into the summarization process. Since the proposed method is an extension of the DILOF framework, the evaluation focuses primarily on methodologically consistent baselines to ensure fair comparison. In addition, hybrid strategies combining count and time triggers are discussed qualitatively due to their unstable behavior observed during the experiments.

V. Results and Discussion

This section presents a comprehensive evaluation of the proposed time-based windowing strategy and analyzes its impact on streaming outlier detection performance under different configurations. To assess the robustness of the proposed framework, the experiments were repeated multiple times under the same parameter settings. The observed AUC values showed low variation across runs, indicating that the proposed method provides stable behavior rather than isolated performance gains. This consistency is particularly important in streaming settings, where irregular arrival patterns may otherwise introduce substantial variability in update behavior and anomaly scores.

A. Performance on Benchmark Datasets

Table III summarizes the AUC values obtained using different time-window configurations across the evaluated datasets. The results demonstrate that the proposed time-based strategy achieves consistently strong and stable performance.

For the Vowel dataset, the AUC values remain high across all configurations, ranging from 0.9428 to 0.9482. The best performance is achieved at a time window of 160 ms, indicating that moderate update intervals provide an effective balance between model responsiveness and stability.

In contrast, the Pendigit dataset exhibits a gradual improvement in performance as the time window increases. The AUC rises from 0.9384 at 100 ms to 0.9637 at 200 ms, suggesting that this dataset benefits from less frequent summarization and longer temporal aggregation.

These findings indicate that while time-based windowing is effective across different datasets, the optimal time window is data-dependent and influenced by the characteristics of the underlying data stream.

B. Comparison with Baseline Methods

To further assess the effectiveness of the proposed approach, we compare it with several baseline methods, including iLOF, MILOF, and the original count-based DILOF.

Fig. 3 and Fig. 4 present the performance comparison across different window configurations for the Vowel and Pendigit datasets, respectively.

On the Vowel dataset, the proposed TD-DILOF consistently achieves the highest AUC values across all window sizes. The iLOF baseline remains nearly constant around 0.90, indicating limited adaptability to changing window configurations. While MILOF [21] and count-based DILOF exhibit gradual improvements as the window size increases, their performance remains noticeably below that of the proposed approach.

Similarly, on the Pendigit dataset, the TD-DILOF demonstrates strong and stable performance across all configurations. Although count-based DILOF achieves competitive results at larger window sizes, it remains slightly inferior to the time-based variant, which consistently maintains higher AUC values.

Notably, the performance advantage of the proposed method is more pronounced at smaller window sizes. This suggests that time-based windowing is particularly effective in scenarios requiring frequent updates, where count-based approaches may suffer from insufficient or irregular sampling.

Overall, these results confirm that incorporating temporal dynamics into the windowing mechanism significantly enhances the robustness and effectiveness of streaming outlier detection compared to traditional count-based methods.

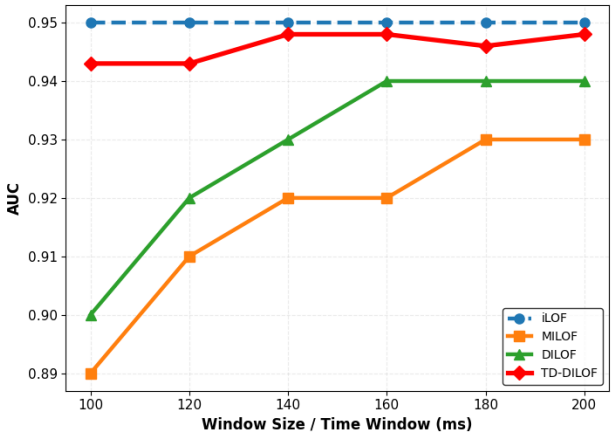


Fig. 3. AUC comparison of the baseline methods and the proposed TD-DILOF on the Vowel dataset.

C. Effect of Time Window on Summarization Behavior

In addition to detection accuracy, we analyze the impact of the time window on summarization behavior and computational cost.

The results show that increasing the time window reduces the number of summarization operations. This behavior is expected, as larger time windows delay the triggering of the summarization process, thereby reducing update frequency.

This behavior highlights a key advantage of time-based windowing, where update frequency is governed by temporal dynamics rather than instance count, enabling more efficient processing in streaming environments.

D. Discussion of Hybrid Strategies

We also investigated hybrid windowing strategies that combine count-based and time-based triggers, including rule-based, score-based, and adaptive formulations.

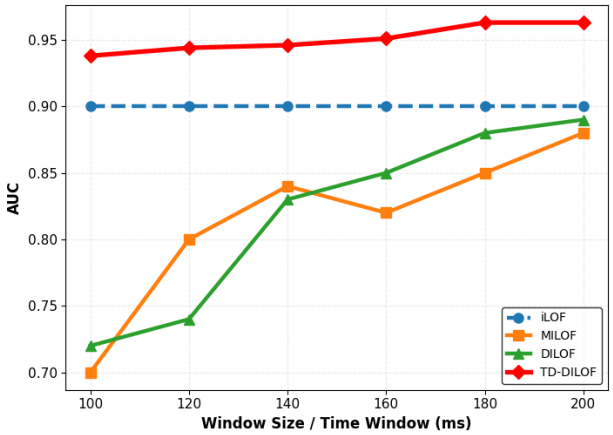


Fig. 4. AUC comparison of the baseline methods and the proposed TD-DILOF on the Pendigit dataset.

However, the experimental results indicate that these hybrid approaches fail to achieve a balanced integration of both mechanisms. In most cases, the hybrid strategy becomes dominated by either the count-based or the time-based trigger, depending on the configuration.

This behavior suggests that simple fusion of count and wall-clock time is insufficient for robust hybridization. More advanced designs, such as event-time-based windowing or adaptive mechanisms that dynamically adjust to stream characteristics, may be required to fully exploit the benefits of hybrid approaches.

Overall, the proposed time-based windowing strategy emerges as the most effective and stable configuration among all evaluated methods, providing a strong balance between accuracy, robustness, and computational efficiency.

TABLE III. AUC Performance Under Different Time-Window Configurations.

Time Window (ms)	Vowel Dataset	Pendigit Dataset
100	0.9429	0.9384
120	0.9428	0.9444
140	0.9481	0.9463
160	**0.9482**	0.9510
180	0.9464	0.9630
200	0.9480	**0.9637**

E. Practical Applications

The proposed framework is particularly suitable for real-world streaming scenarios in which data arrival rates are naturally irregular and bursty. Examples include IoT sensor monitoring, industrial fault detection, network intrusion detection, financial transaction surveillance, and edge-based anomaly detection. In such environments, time-driven summarization provides a more realistic control mechanism than fixed count-based updates, allowing the model to maintain temporal consistency while avoiding unnecessary computational overhead.

VI. Limitations

Despite the promising results, this study has several limitations. First, the experiments were conducted in a simulated streaming environment that approximates real-time data processing conditions. Although this setup enables controlled evaluation of the proposed framework, further validation in fully real-time deployment scenarios would provide additional practical insights.

Second, the time measurement used in this study is based on system processing time rather than event time. Consequently, the behavior of the proposed method may vary across hardware settings and deployment environments.

Third, the hybrid windowing strategies explored in this work rely on relatively simple formulations. These approaches were not sufficient to achieve a balanced integration between count-based and time-based triggers, indicating the need for more advanced designs.

These limitations open promising directions for future research, particularly for real-time deployment, event-time modeling, and evaluation on larger and more diverse streaming datasets.

VII. Conclusion

This study demonstrates that temporal awareness is a critical yet underexplored factor in streaming outlier detection. Using Density Incremental Local Outlier Factor (DILOF) as a baseline framework, we investigated a time-driven windowing strategy in which model summarization is triggered by elapsed time rather than by a fixed number of arriving instances.

The experimental results show that the proposed TD-DILOF framework consistently achieves strong and stable detection performance across different benchmark datasets and time-window configurations. The method produces high AUC values while also revealing that the optimal time window depends on the characteristics of the underlying data stream. This finding emphasizes that update triggering in streaming systems should account for temporal behavior rather than relying solely on instance count.

From a computational perspective, the proposed framework provides an improved balance between detection accuracy and efficiency by reducing unnecessary summarization operations under varying arrival rates. In addition, the analysis of hybrid strategies shows that simple combinations of count-based and time-based triggers are often dominated by one mechanism, limiting their practical effectiveness.

Overall, the findings suggest that time-driven windowing is not only beneficial but also practically meaningful for robust streaming outlier detection, particularly in real-world environments such as IoT monitoring, industrial systems, network security, and edge computing, where data arrival behavior is inherently dynamic and temporally irregular.

VIII. Future Work

Several directions can extend the present work. First, adaptive time window selection strategies should be investigated to dynamically adjust the window size according to stream

characteristics such as arrival rate, density variation, and concept drift. Such strategies may reduce sensitivity to dataset-specific configurations and improve robustness in practical deployments.

Second, more principled hybrid windowing mechanisms can be developed to integrate count-based and time-based information without suffering from trigger dominance. This may require dynamic control rules, event-time formulations, or learning-based update policies.

Third, replacing system-time processing with event-time-based windowing represents a promising direction, especially in distributed and real-time environments where temporal consistency is essential.

Finally, future studies should validate the proposed framework on a wider range of benchmark and real-world datasets, and compare it with additional state-of-the-art streaming anomaly detection methods to further assess scalability, generalizability, and deployment readiness.

REFERENCES

[1] M. Fragkoulis, P. Carbone, V. Kalavri, and A. Katsifodimos, "A survey on the evolution of stream processing systems," *The VLDB Journal*, 2024.

[2] S. Anjum, S. K. Yadav, and S. Yadav, "Stream data model and architecture," in *Machine Learning: Navigating the Big Data Era*, Springer, 2024.

[3] R. Hu and Y. Wang, "An efficient outlier detection algorithm for data streaming," *arXiv preprint arXiv:2501.01061*, Jan. 2025.

[4] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 4, no. 3, p. 20, 2020.

[5] M. Sabha and B. Tugrul, "Stream data analysis and processing frameworks for detecting outliers in human activities: A review," *WSEAS Transactions on Information Science and Applications*, 2025.

[6] Z. Hu *et al.*, "ASOD: An adaptive stream outlier detection method using online strategy," *Journal of Cloud Computing*, 2024.

[7] I. Altaf and M. A. Chachoo, "Advances in density-based outlier detection algorithms: Exploration of LOF with experimental analysis," *Applied Soft Computing*, May 2025.

[8] L. Tran, L. Fan, and C. Shahab, "Distance-based outlier detection in data streams," *Proceedings of the VLDB Endowment*, vol. 9, no. 12, 2016.

[9] F. Angiulli and F. Fassetti, "Detecting distance-based outliers in streams of data," in *Proc. ACM CIKM*, 2007, pp. 811–820.

[10] L. Tran, M. Y. Mun, and C. Shahabi, "Real-time distance-based outlier detection in data streams," 2021.

[11] L. Cao *et al.*, "Scalable distance-based outlier detection over high-volume data streams," in *Proc. IEEE ICDE*, 2014, pp. 76–87.

[12] A. Gounaris and Y. Manolopoulos, "Continuous monitoring of distance-based outliers over data streams," in *Proc. IEEE ICDE*, 2011, pp. 135–146.

[13] G. S. Na, D. Kim, and H. Yu, "DILOF: Effective and memory efficient local outlier detection in data streams," in *Proc. ACM SIGKDD*, 2018.

[14] F. Bonomi *et al.*, "Fog computing and its role in the IoT," in *Proc. MCC Workshop*, 2012.

[15] A. Ahmed and S. H. Mohammad, "A survey on fog computing: Architecture, key technologies, applications and open issues," *Journal of Network and Computer Applications*, vol. 98, pp. 27–42, 2017.

[16] S. Yi *et al.*, "Fog computing: Platform and applications," in *Proc. IEEE HotWeb*, 2015, pp. 73–78.

[17] S. Khan *et al.*, "A comprehensive survey on multifaceted fog computing resource management techniques, trends, applications and future directions," 2025.

[18] M. M. Breunig *et al.*, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD*, 2000.

[19] M. Hassaan *et al.*, "A fast and efficient algorithm for outlier detection over data streams," *International Journal of Advanced Computer Science and Applications*, 2021.

[20] D. Pokrajac *et al.*, "Incremental local outlier detection for data streams," in *IEEE CIDM*, 2007, pp. 504–515.

[21] M. Salehi *et al.*, "Fast memory efficient local outlier detection in data streams," *IEEE Transactions on Knowledge and Data Engineering*, 2016.

[22] A. Nilofer and S. Sasikala, "Enhanced hybrid filter based feature selection models for high dimensional streaming data," *IAENG International Journal of Computer Science*, vol. 52, no. 12, pp. 4678–4691, 2025.

[23] R. Hu *et al.*, "EILOF: An efficient incremental local outlier factor algorithm for data streaming," *arXiv preprint arXiv:2501.01061*, 2025.

[24] K. Moeenfar *et al.*, "EiForestASD: A fast and accurate tree-based approach for anomaly detection in streaming data," *International Journal of Data Science and Analytics*, 2025.

[25] A Duraj *et al.*, " Detection of anomalies in data streams using the LSTM-CNN model" *Sensors, mdpi*, 2025.

[26] QT Tran *et al.*, "Concept drift detection in image data stream: a survey on current literature, limitations and future directions" *Artificial Intelligence Review, Springer*, 2026.

Deep Learning-Based Road Damage Detection Using Improved YOLOv8: Model Performance and Implementation

Aulia Rahman[1], Hwa Jen Yap[2], Rusdha Muharar[3]
Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia[1,3]
Faculty of Engineering-Department of Mechanical Engineering, Universiti Malaya, Kuala Lumpur, Malaysia[2]

***Abstract*—Road infrastructure monitoring plays a crucial role in ensuring safety and economic efficiency; however, conventional manual inspection methods are expensive and resource-intensive. This study presents the development and evaluation of a real-time road damage detection system using the improved YOLOv8 architecture. Building upon the strengths of previous YOLO models, YOLOv8 offers enhanced accuracy and inference speed, making it highly suitable for mobile deployment. The model was trained to identify six common road damage types in Indonesia: potholes, alligator cracks, transverse cracks, longitudinal cracks, edge cracks, and road joints. Utilizing a hybrid dataset of 2,946 images, combining locally collected data and the RDD2020 dataset, the system incorporates mosaic augmentation and optimized preprocessing to improve generalization. The optimized YOLOv8 model achieved a mean Average Precision (mAP@50) of 96.3%, an F1-score of 91%, and an overall accuracy of 91%, demonstrating superior detection and classification performance. The system was deployed as a user-friendly smartphone application, enabling automated, geo-tagged road condition surveys and offering road authorities a scalable, efficient, and practical tool for infrastructure monitoring.**

***Keywords*—Road damage detection; YOLOv8; deep learning; computer vision; mobile application**

I. Introduction

Road infrastructure is a crucial public asset that requires regular monitoring because of its significant role in fostering economic development [1], [2]. However, road surfaces inevitably deteriorate over time because of both natural and human influences. In Indonesia, motorists often encounter damaged roads resulting from extreme weather, heavy traffic loads, substandard construction quality, and insufficient maintenance. These conditions frequently lead to accidents, some of which are fatal, as drivers struggle to anticipate and navigate road defects. Damaged roads also disrupt transportation, increasing vehicle operating costs because of damage from potholes and other defects [3]. To mitigate accident risks and enhance transportation efficiency, monitoring road conditions is essential [4].

Currently, road damage surveys are conducted manually by physically marking damaged areas by inspectors. However, this approach is labor-intensive, time-consuming, and impractical for extensive road networks [5]. Continuous monitoring is challenging due to the vastness scope of road infrastructure. To address this, automated solutions utilizing computer vision and deep learning are increasingly being employed for road damage monitoring [6], [7], [8].

Previous studies have demonstrated the use of YOLO models for road damage detection, as seen in [9], [10], [11]. More recently, YOLOv5, using a 640×640 image size, achieved an F1 score of 66.51% with the RDD2022 dataset, detecting potholes, alligator cracks, transverse cracks, and longitudinal cracks [12].

Although previous studies have demonstrated the potential of YOLO-based approaches for road defect detection, a significant research gap remains in the development of models adapted to local road conditions and a wider variety of defect categories. Many publicly available datasets do not include classes such as edge cracks and road joints, despite these defects being common and important indicators of pavement condition in countries like Indonesia.

This study aims to fill the existing gap by developing and implementing a robust road damage detection system using improved YOLOv8. The primary goal was to create a model that not only accurately identifies common cracks and potholes but also detects two additional categories—edge cracks and road joints—using a novel dataset from Indonesian roads. The key scientific contributions of this study are as follows: 1) the creation of a comprehensive and locally-relevant dataset for road damage; 2) the development of a highly accurate YOLOv8 detection model, achieving an F1-score of 91% F1-score, which surpasses previous benchmarks through tailored data processing and augmentation; and 3) the successful integration of this model into an on-device smartphone application, offering a practical solution for automated infrastructure monitoring.

II. Related Works

Object detection is a branch of computer vision that focuses on locating and identifying objects within images or videos. The fundamental components of object detection consist of three main elements: bounding box, class label, and confidence score [13], [14]. Bounding boxes are utilized to pinpoint the positions of objects in an image. The class label specifies the type of object detected, whereas the confidence score reflects the certainty level that the detected object corresponds to the predicted class [15].

Recent advancements in deep learning have significantly enhanced the performance of object detection models. Two notable architectures, the Faster R-CNN [16] and YOLO (You Only Look Once) [17], exemplify key approaches to object detection. The Faster R-CNN employs a two-stage process:

generating region proposals followed by classification, whereas YOLO is a single-stage detector that predicts bounding boxes and class labels in one forward pass, making it ideal for real-time applications. Although two-stage detectors can achieve high accuracy, their slower inference speeds render them less suitable for on-the-fly detection from a moving vehicle. These methods greatly enhance the speed and accuracy of object detection tasks, offering powerful tools for various applications, including autonomous driving and surveillance. The next section discusses the building blocks of object detection used in this research: convolutional neural networks (CNNs) and YOLOv8.

A. Convolutional Neural Network

Convolutional neural networks (CNNs) are a type of artificial neural network that are frequently employed for processing image data [18], [19]. At the heart of CNNs is the convolution operation, which is essential to their operation and functions as a filter. These networks are composed of multiple layers that are specifically crafted to perform convolutions on image data. The training process of the model encompasses three primary stages: the convolutional layer, the pooling layer, and the fully connected layer. CNNs have become prominent because of their capacity to effectively capture spatial hierarchies within image data. Their architecture mirrors the structure of the human visual cortex, enabling them to focus on localized regions of an image through convolutional filters. As the network deepens, these filters progressively learn more intricate features, allowing CNNs to identify patterns, such as edges and textures, in the initial layers and more abstract patterns in the deeper layers [20]. Pooling layers reduce the dimensionality of the feature maps while preserving crucial information and minimizing computational costs. This makes CNNs highly effective for various image-based tasks, including classification, object detection, and segmentation [21].

Since AlexNet's success in the ImageNet competition in 2012, CNN architectures have undergone significant evolution. Models such as VGG [22], ResNet [23], and Inception [24] have enhanced both accuracy and computational efficiency. Furthermore, the adoption of transfer learning has broadened CNN applications, enabling models pretrained on extensive datasets to be fine-tuned for specific tasks. This approach boosts model performance, especially in situations where labeled data are scarce [25], [26]. The robustness and versatility of CNNs have made them indispensable not only in computer vision but also in other fields such as medical imaging, where they support disease diagnosis through the analysis of radiological data [27] and thermal images [28].

B. YOLOv8

YOLO, short for "You Only Look Once", has transformed object detection by approaching it as a single regression problem, thereby moving away from traditional sliding window techniques. By processing the entire image in one go, YOLO predicts bounding boxes and class probabilities in real-time, making it ideal for high-speed applications such as autonomous driving and surveillance [29]. Building on this foundation, YOLOv5 features a lightweight yet robust architecture that utilizes a CSPNet-based backbone for feature extraction and a path aggregation network (PANet) to enhance multi-scale

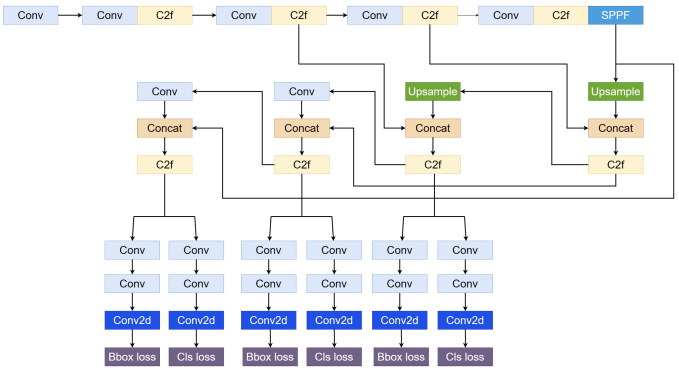


Fig. 1. Original YOLOv8 architecture

object detection. This combination of speed and accuracy has made YOLOv5 a favored choice for real-time object detection tasks across various domains. An application of YOLOv5 in real-time vehicle detection was conducted [30].

YOLOv8 is an advanced object detection architecture designed for real-time applications, which combines high inference speed with enhanced accuracy [31], [32]. It builds upon the YOLO series by utilizing a cross-stage partial network (CSPNet) backbone for efficient feature extraction, a neck component that refines features through additional convolutional layers, and a detection head that incorporates a feature pyramid network (FPN) to detect objects at multiple scales. This architecture enables precise detection and classification of objects, making YOLOv8 particularly suitable for tasks that require fast and accurate detection, such as road damage monitoring on mobile platforms.

This study introduces a road damage detection and classification system utilizing the YOLOv8 architecture, as shown in Fig. 1. This system was specifically designed to identify six common types of road damage found on Indonesian roads: potholes, alligator cracks, transverse cracks, longitudinal cracks, edge cracks, and road joints. The YOLOv8 architecture was selected for its exceptional balance between real-time inference speed and high detection accuracy, making it an ideal choice for deployment on a mobile platform for on-road monitoring. By leveraging the YOLOv8 model, this system aims to deliver accurate and efficient detection tailored to local road conditions.

C. Bidirectional Feature Pyramid Network (BiFPN)

The Bidirectional Feature Pyramid Network (BiFPN) represents a significant advancement in the design of the neck component for modern object detection architectures, particularly in improving efficient multi-scale feature fusion. Introduced as part of the EfficientDet family [33], BiFPN addresses the limitations of conventional Feature Pyramid Networks (FPN) that rely on static fusion strategies. It incorporates bidirectional connections (top-down and bottom-up) along with learnable weighting mechanisms, enabling the model to adaptively determine the relative importance of each input feature map. This approach effectively mitigates feature imbalance across scales and enhances detection performance, especially for small objects, while maintaining computational efficiency [34], [35].

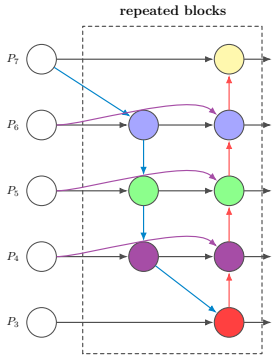


Fig. 2. BiFPN feature network design.

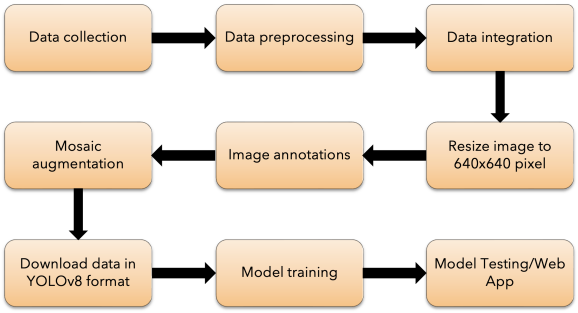


Fig. 3. Road damage detection and classification workflow.

From an architectural perspective, BiFPN operates by taking multiple feature maps from the backbone (e.g., levels P2–P5) and fusing them through nodes equipped with learnable, non-negative weights. These weights are activated using a ReLU function and normalized to form a convex combination, ensuring stable training while allowing the network to suppress less informative features. Additionally, BiFPN is designed with computational efficiency in mind by minimizing redundant connections and limiting the number of inputs per fusion node, thereby reducing FLOPs compared to traditional multi-scale fusion methods. The fusion process can also be iteratively repeated to progressively refine feature representations, which is particularly beneficial in scenarios with significant scale variation, such as UAV imagery or micro-defect detection [36] and surface flaw detection of a small automotive component [37]. This balance between efficiency and representational power has led to the widespread adoption and further development of BiFPN across various object detection frameworks. The BiFPN feature network design is shown in Fig. 2.

In this study, the original feature aggregation layers in the YOLOv8 neck were modified by replacing the conventional PAN-FPN structure with a BiFPN module, enabling more efficient fusion of low-level spatial features and high-level semantic features. This architectural modification improves the model's capability to detect road defects of varying sizes, particularly small and irregular crack patterns. Although the addition of BiFPN increases the number of feature fusion operations and introduces a slight computational overhead, the impact on inference speed remains manageable due to the lightweight design and optimized weighted connections of the BiFPN architecture.

D. Evaluation Metrics

Model performance is assessed using the Intersection over Union (IoU) score to measure bounding box accuracy and the mean Average Precision (mAP) to evaluate object detection accuracy [38]. The IoU score is determined by calculating the ratio of the overlap area between the predicted bounding box and the ground truth bounding box to the union area of these boxes. A higher IoU score indicates a closer alignment between the predicted and actual object locations.

The recognition accuracy was assessed using mAP, which integrates both precision and recall across various IoU thresholds, thereby offering a thorough evaluation of the model's performance under different conditions [39].

As illustrated in Fig. 3, this work comprises eight main steps: data collection, data cleaning, data integration, image resizing, image annotation, applying mosaic augmentation, downloading the dataset in the YOLOv8 PyTorch format, training the model, model testing, and application implementation.

E. Data Collection

The dataset for this study was independently gathered using a smartphone camera. To ensure high-quality images, the camera was stabilized by minimizing body movement and focusing on road damage under optimal lighting conditions. Data collection occurred along a national road in Aceh province in Indonesia, capturing various types of road damage. The dataset encompasses six categories of road damage: potholes, alligator cracks, transverse cracks, longitudinal cracks, edge cracks, and road joints. To balance the dataset, additional road images were sourced from the RDD2020 dataset, particularly from Japanese roads. The final dataset comprised 2,946 images, divided into 2,049 images for training, 600 for validation, and 294 for testing. The total number of images for each class is detailed in Table I.

TABLE I. ADDITIONAL LOCAL DATASET

Class	Amount
Pothole	451 images
Alligator crack	368 images
Longitudinal crack	327 images
Transverse crack	432 images
Edge crack	491 images
Road joints	491 images

F. Data Prepocessing

Dataset preprocessing, including data cleaning and data integration, is an essential step to ensure that the dataset used for analysis or model development is accurate, consistent, and complete. Its primary objective is to detect and correct errors, remove inconsistencies, and handle missing values. By improving the quality of the data, this process enhances the precision and reliability of the analysis or model, ultimately leading to more accurate and impactful results.

Following the completion of the data cleaning process, the subsequent step involves merging the independently collected data by the researchers with the RDD2020 dataset from Japan.

The combined dataset, which is crucial for the analysis, is presented in Table II.

TABLE II. RESULT OF MERGING DATASETS

Class	Independent dataset	RDD2020	Amount
Pothole	451 images	40 image	491 images
Alligator crack	368 images	123 images	491 images
Longitudinal crack	327 images	164 images	491 images
Transverse crack	432 images	59 images	491 images
Edge crack	491 images	-	491 images
Road joints	491 images	-	491 images
Total			2946 images

The subsequent step involves resizing all images to 640x640 pixels. This specific dimension was selected to offer a higher resolution, thereby enabling the model to capture more detailed features of objects, which is essential for enhancing detection accuracy, particularly for smaller objects. Moreover, employing larger image sizes during training allows the model to learn more intricate features, thereby improving its overall performance.

G. Image Annotation

The data annotation process is designed to deliver precise information regarding the location and category of objects within images by marking bounding boxes around them. During this stage, the dataset is imported into Roboflow, where the bounding box tool is employed to accurately outline the objects, as illustrated in Fig. 4.

Each object is encased in a polygon that outlines its shape, with class labels assigned based on the object type. The deep learning model then learns to identify objects by examining the patterns and features within these bounding boxes. Detailed and precise annotations greatly enhance the quality of the training data, thereby improving the model's capacity to detect and classify road damage. By ensuring that each object is accurately labeled and represented, the model is better prepared to generalize from the training data and perform effectively in real-world scenarios.

H. Data Augmentation

Data augmentation includes a variety of techniques designed to expand and enrich datasets for machine learning models, with mosaic augmentation being one of the most widely used approaches. The process begins with an input image, after which the mosaic technique applies several transformations, including flipping, rotation, shearing, brightness adjustment, contrast enhancement, and noise injection, before combining regions from multiple images into a single composite sample [40], [41]. The process of our mosaic augmentation procedure is illustrated in Fig. 5.

This technique is widely employed in training object detection models, particularly for computer vision tasks. Mosaic augmentation merges multiple images into a single composite training sample, thereby enabling the model to learn from a broader range of examples [42], [43], [44]. This approach was selected over simpler geometric augmentations, such as flips and rotations, to significantly boost model robustness. By combining four different images into a single training sample, mosaic augmentation exposes the model to a wider variety of object scales, backgrounds, and partial occlusions, thereby improving its ability to accurately detect small or partially hidden defects on complex road surfaces.

I. Experimental Environment

The experimental environment consisted of a Windows 10 operating system, equipped with an NVIDIA GeForce RTX 3090 graphics card featuring 24 GB of GDDR6X memory, an Intel i7-10700 CPU, and 32 GB of RAM. Before initiating the training process, it is crucial to meticulously adjust the hyperparameters to ensure that the model learns effectively and yields optimal results. The key settings included the number of epochs, batch size, optimizer, momentum, and learning rate, all of which were customized to suit the data and model characteristics.

This research develops a deep learning model using the YOLOv8 architecture to detect six types of road defects: potholes, alligator cracks, transverse cracks, longitudinal cracks, edge cracks, and road joints. This section explains the process of building the baseline model and refining it to optimize the hyperparameters, ultimately creating a robust model for detecting and classifying these road damage types. The process involves fine-tuning settings such as the number of epochs, batch size, optimizer, momentum, and learning rate, all of which are adjusted to align with the characteristics of both the dataset and the model. Furthermore, robustness tests were conducted to evaluate the model's ability to generalize across

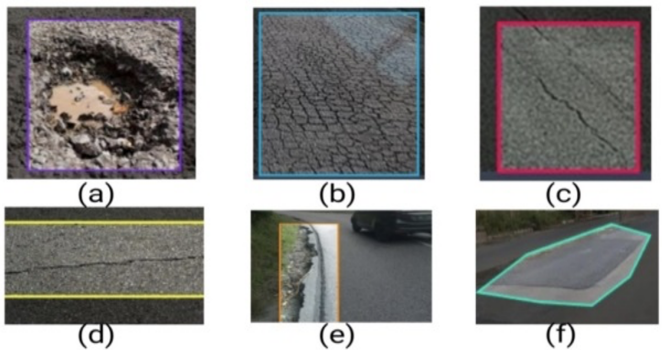


Fig. 4. Image annotation process: Creating a bounding box to enclose road damage.

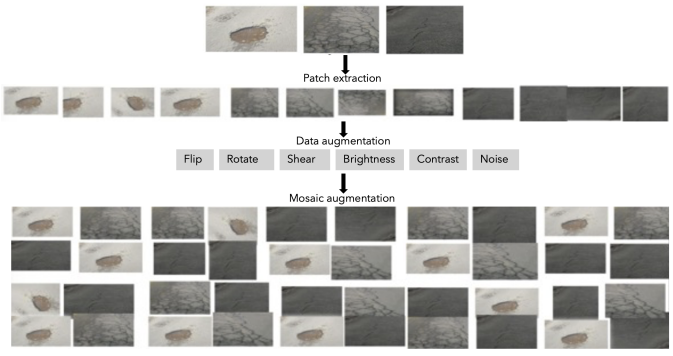


Fig. 5. Mosaic augmentation procedure.

different environmental conditions, such as varying lighting, road textures, and occlusions. The hyperparameters used in model training are shown in Table III.

TABLE III. HYPERPARAMETERS USED IN THE TRAINING PROCESS

Hyperparameter	YOLOv8
Epoch	200
Batch size	16
Optimizer	SGD
Learning rate	10^{-2}
Momentum	0.9
Image size	640x640

III. EXPERIMENTAL SETUP AND RESULTS

A. Training Results and Robustness Analysis

To evaluate the performance of the trained model, various metrics were examined, including loss curves, precision-recall metrics, confusion matrices, and mean Average Precision (mAP) scores. As shown in Fig. 6, the training loss curve starts at a high value but decreases significantly early in the training process. With an increasing number of epochs, the training loss continues to decline, indicating that the model is enhancing its predictions on the training data without significant overfitting. In contrast, the validation loss curve begins at a lower value but rises around epoch 75, likely due to variations in road damage patterns. However, with further training, the validation loss stabilizes and then decreases, suggesting that the model generalizes well to unseen data. The training of the model is illustrated in Fig. 6.

To further confirm the effectiveness of the model, we performed additional robustness tests by assessing its performance under various road surfaces and environmental conditions. Preliminary qualitative observations suggest that the model remains reasonably stable under moderate lighting variations and different road textures; however, comprehensive robustness testing under adverse environmental conditions remains future work.

Fig. 7 illustrates the precision and recall metric curves observed during training. Utilizing a learning rate of 10^{-2} alongside a mosaic augmentation strategy enabled the model to achieve an optimal balance between precision and recall. Both precision and recall stabilize around 0.9, demonstrating

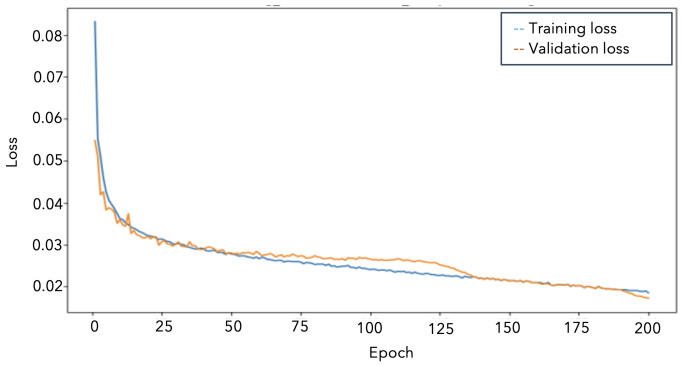


Fig. 6. Training and validation loss of the model.

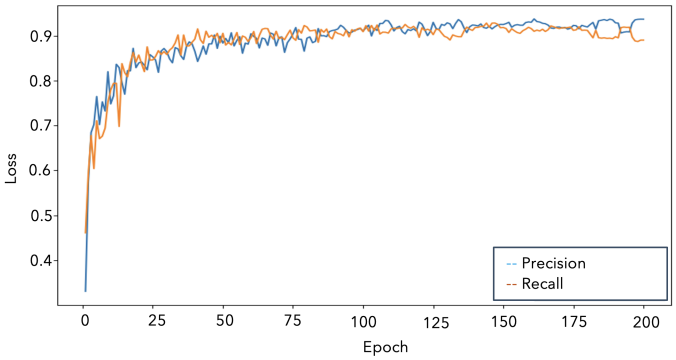


Fig. 7. Precision and recall of the model.

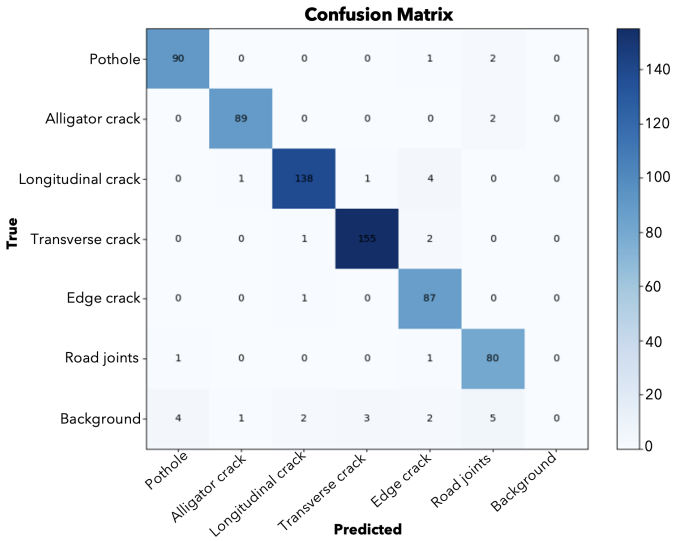


Fig. 8. Confusion matrix of model prediction.

the model's capability to accurately detect road damage while minimizing false positives.

The confusion matrix analysis, depicted in Fig. 8, illustrates the testing results using 294 enhanced images via mosaic techniques. The findings indicate that within the pothole class, 90 instances were accurately identified as potholes, while one was misclassified as an edge crack and two as road joints. In the alligator crack category, 89 were correctly recognized; however, two were misclassified as road joints. For the longitudinal crack class, 138 were correctly detected; one was misclassified as a transverse crack, and four transverse cracks were incorrectly identified as edge cracks. In the edge crack class, 87 instances were correctly classified; one was misclassified as a longitudinal crack. Regarding the road joint class, 80 predictions were accurate, while one was misclassified as a pothole and one as an edge crack. The confusion matrix underscores a significant improvement, as most predictions are correct, with only a few classification errors, demonstrating the model's robust performance.

Fig. 9 presents the F1-score results from the training process. The x-axis indicates the model's confidence level in its predictions, with higher confidence levels positioned on the right. The y-axis displays the F1 score, a metric that balances

precision and recall. The blue line (pothole) performs well, maintaining a high F1-score across most confidence levels. The orange line (transverse crack) begins with a high F1-score but drops significantly as confidence increases, likely owing to prediction errors, as the transverse crack class closely resembles the longitudinal crack class. The green line (longitudinal crack) demonstrates stable and consistent performance throughout. The red line (edge crack) also shows strong performance, consistently achieving a higher F1-score than the other classes.

The purple line, representing alligator cracks, demonstrates stable and robust performance. Similarly, the brown line, which denotes road joints, performs well, similar to the pothole class. The thick blue line illustrates the combined F1 score for all classes. At a confidence level of 0.556, this combined F1 score reaches 0.91, indicating excellent overall model performance.

Table IV presents the mAP scores for each class. At 0.5, the mAP value is outstanding, with the edge crack class achieving the highest score of 0.980, indicating the model's exceptional performance in detecting road damage with minimal error. This result indicates strong localization and classification performance at the IoU threshold of 0.5. with the ground truth box, showcasing high accuracy at this threshold. Furthermore, the mAP value at 0.5:0.95 also yields good results, with the edge crack class reaching a peak value of 0.698. Although performance slightly diminishes at higher IoU thresholds, the model continues to exhibit strong detection capabilities. The modest decline from mAP 0.5 to mAP 0.5:0.95 underscores the model's ability to manage tighter bounding box predictions, maintaining a balance between accuracy and precision.

TABLE IV. THE MAP VALUE OF EACH CLASS

Class	*Precision*	*Recall*	*mAP 0,5*	*mAP 0.5:0.95*
Pothole	0.916	0.941	0.976	0.676
Alligator crack	0.77	0.862	0.91	0.64
Transverse crack	0.702	0.683	0.708	0.589
Longitudinal crack	0.837	0.828	0.88	0.674
Edge crack	0.92	0.912	0.98	0.698
Road joints	0.734	0.805	0.921	0.692

We conducted the same experiment with the hyperparameters outlined in Table III to benchmark our proposed model against YOLOv3, YOLOv4, and YOLOv5. The performance of each model is presented in Table V, which demonstrates that YOLOv8 BiFPN exhibits superior performance.

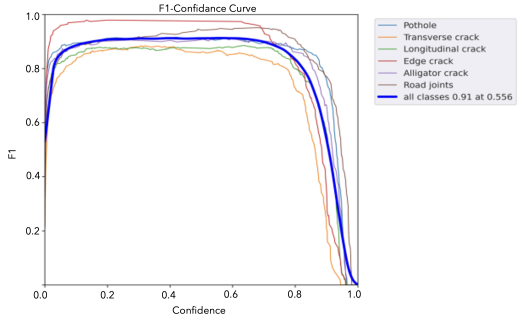
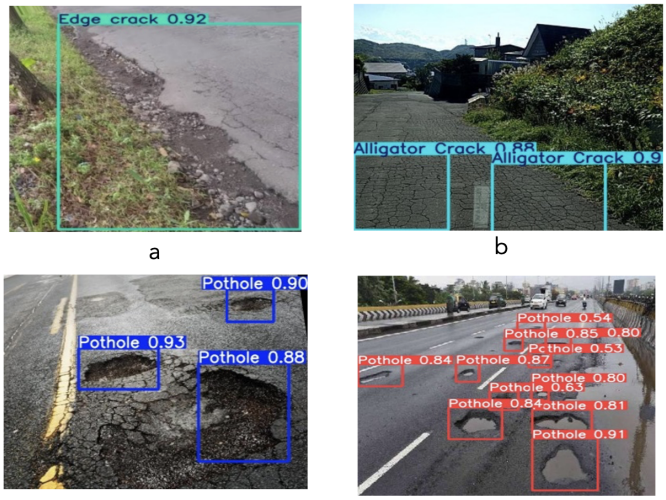


Fig. 9. F1-score of the model.



a b c d

Fig. 10. Road damage detection test: a) Edge crack, b) Alligator crack, c) Pothole, d) Multiple potholes.

TABLE V. PERFORMANCE COMPARISON AGAINST PREVIOUS YOLO ARCHITECTURES.

MODEL	Precision	Recall	mAP@0.5
YOLOv8 BiFPN	0.891	0.905	0.963
YOLOv8	0.873	0.893	0.955
YOLOv5	0.813	0.839	0.896
YOLOv4	0.793	0.818	0.88
YOLOV3	0.632	0.702	0.734

B. Detection Model Evaluation

Once the model for road damage detection and classification is trained, the subsequent step involves testing to evaluate its performance. This test determines how effectively the model can identify and categorize various types of road damage in previously unseen data. Additionally, it assesses the model's accuracy, precision, and recall across different predefined road damage classes. The results of the model testing are shown in Fig. 10.

C. Model Implementation on Smartphone

We developed a deep learning-based application utilizing the YOLOv8 architecture to detect and classify road defects into six primary categories. The trained model can accurately identify various defect types, including potholes, transverse cracks, longitudinal cracks, edge cracks, alligator cracks, and road joints. Owing to its real-time inference speed and high accuracy, this application enables road managers to perform automated inspections, thereby accelerating the damage detection process and enhancing the efficiency of road repair efforts.

The device was installed on the dashboard of a public passenger car. Our developed application captures images of the road surface approximately 10 m ahead, showcasing its capability to capture images while in motion without any data loss or duplication at an average speed of 40 km/h

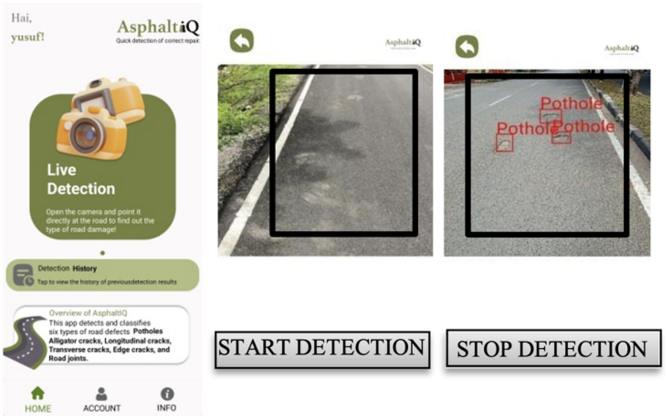


Fig. 11. On-device detection application on a smartphone.

(approximately 10 m/s), with one image captured per second. Furthermore, the application can accurately detect road defects within 1.5 s. A screenshot of the real-time detection application on a smartphone is shown in Fig. 11.

The operating screen of the smartphone application was mounted on the dashboard of a passenger car. The detection of road surface damage commenced when the "START DETECTION" button was activated. Images of the detected damage, along with their position information, were transmitted to the detection history in the database. Only images of confirmed damage were retained in the detection history.

The application features an image upload capability that enables users to submit photos of damaged roads. It automatically identifies and categorizes the type of damage depicted in an image. Powered by a trained YOLOv8 model, the application delivers fast and accurate detection results, simplifying the process for users to evaluate road conditions directly from uploaded images.

IV. Discussion

The performance of our YOLOv8 BiFPN model demonstrated a marked improvement over previous studies cited in the literature. For example, our model achieved an F1-score of 0.91, significantly surpassing the 50.39% reported by Opara [9] using YOLOv3 and the 66.51% using YOLOv5 on the RDD2022 dataset [12]. This significant improvement in performance can be attributed to several factors: the use of a tailored dataset that includes specific local road damage types (edge cracks and road joints), a larger and more balanced dataset, and the effective application of mosaic augmentation, which enhanced the model's ability to learn diverse features. The high mAP of 98% (at IoU=0.5) further confirms the model's accuracy in correctly detecting and classifying defects. The high mAP@50 should be interpreted as performance on the current test distribution rather than as universal generalization. The relatively small test set and visual similarity among locally collected road images may have contributed to the high score. Therefore, mAP@50:95 and future cross-region validation are reported/discussed to provide a more conservative assessment.

Despite these results, the study has several limitations. The performance of the model is highly dependent on environmental conditions, and its accuracy may decline in adverse weather, such as heavy rain or fog, or at night with insufficient street lighting. The training data were primarily collected from national roads in Indonesia, which may introduce a geographic bias. Consequently, the model may not perform as well on different types of road surfaces (e.g., concrete or unpaved roads) without additional training data. Furthermore, the tests were conducted at a moderate speed (30 km/h), and performance at significantly higher speeds has not yet been tested.

Because the local dataset was collected from a national road in Indonesia, the model may learn region-specific pavement textures, lane markings, lighting conditions, and defect appearances. Therefore, the current model should be regarded as optimized for the studied Local road context, and additional cross-region validation is required before wider deployment.

The proposed system demonstrated promising performance under daytime conditions; however, its robustness under heavy rain, fog, and nighttime illumination has not yet been comprehensively validated. Therefore, claims of real-world deployment suitability are limited to conditions similar to those represented in the test data.

The practical implications of this research are substantial. The successful implementation of the model into a smartphone application provides road management authorities with a practical and cost-effective tool for road infrastructure monitoring. This system can replace slow and labor-intensive manual inspections with rapid automated surveys. This enables authorities to quickly identify critical areas, prioritize repairs based on the severity and type of damage, and allocate maintenance budgets more efficiently.

V. Conclusion and Future Work

We successfully developed two additional dataset classes, namely, edge cracks and road joints, and evaluated the performance of the YOLOv8 architecture on these new classes. For the edge crack class, using a learning rate of 10^{-2}, the mAP value at 0.5 reached 0.963, demonstrating high prediction accuracy. A well-performing model was developed using a modified YOLOv8 architecture, achieving an mAP 0.5 score of 96.3% and an F1 score of 91% with mosaic augmentation preprocessing, indicating excellent performance in detecting road damage.

An application was successfully developed using the YOLOv8 architecture to detect and classify road damage into six categories: potholes, transverse cracks, longitudinal cracks, edge cracks, alligator cracks, and road joints. This application enables users to upload images of road damage for on-device detection and classification, offering a practical tool for efficient road maintenance and inspection.

Although the model demonstrated high accuracy under controlled daytime conditions, future work is necessary to assess its performance in more challenging environments. Future studies may include testing the system under nighttime conditions, where limited illumination and shadows could affect detection accuracy. In addition, the model should be evaluated on different road surface materials, such as concrete roads and unpaved roads, to ensure broader applicability across

diverse infrastructure conditions. Expanding the dataset to include these scenarios is expected to improve the adaptability and reliability of the YOLOv8-based detection system for real-world deployment.

REFERENCES

[1] Y. Song, P. Wu, Q. Li, Y. Liu, and L. Karunaratne, "Hybrid Nonlinear and Machine Learning Methods for Analyzing Factors Influencing the Performance of Large-Scale Transport Infrastructure," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 287–12 300, Aug. 2022.

[2] Y. Song, P. Wu, D. Gilmore, and Q. Li, "A Spatial Heterogeneity-Based Segmentation Model for Analyzing Road Deterioration Network Data in Multi-Scale Infrastructure Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7073–7083, Jun. 2020.

[3] M. Emu, F. B. Kamal, S. Choudhury, and Q. A. Rahman, "Fatality Prediction for Motor Vehicle Collisions: Mining Big Data Using Deep Learning and Ensemble Methods," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 199–209, Jan. 2022.

[4] H. Xin, Y. Ye, X. Na, H. Hu, G. Wang, C. Wu, and S. Hu, "Sustainable Road Pothole Detection: A Crowdsourcing Based Multi-Sensors Fusion Approach," *Sustainability*, vol. 15, no. 8, p. 6610, Apr. 2023.

[5] J. Wang and T. Ueda, "A review study on unmanned aerial vehicle and mobile robot technologies on damage inspection of reinforced concrete structures," *Structural Concrete*, vol. 24, no. 1, pp. 536–562, Jan. 2023.

[6] A.-P. Botezatu, A. Burlacu, and C. Orhei, "A Review of Deep Learning Advancements in Road Analysis for Autonomous Driving," *Applied Sciences*, vol. 14, no. 11, p. 4705, May 2024.

[7] F. Wan, C. Sun, H. He, G. Lei, L. Xu, and T. Xiao, "YOLO-LRDD: a lightweight method for road damage detection based on improved YOLOv5s," *EURASIP Journal on Advances in Signal Processing*, vol. 2022, no. 1, Oct. 2022.

[8] S.-S. Park, V.-T. Tran, and D.-E. Lee, "Application of Various YOLO Models for Computer Vision-Based Real-Time Pothole Detection," *Applied Sciences*, vol. 11, no. 23, p. 11229, Nov. 2021.

[9] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 12, pp. 1127–1141, Jun. 2018.

[10] L. A. Silva, V. R. Q. Leithardt, V. F. L. Batista, G. Villarrubia González, and J. F. De Paz Santana, "Automated Road Damage Detection Using UAV Images and Deep Learning Techniques," *IEEE Access*, vol. 11, pp. 62 918–62 931, Jan. 2023.

[11] Y. Li, C. Yin, Y. Lei, J. Zhang, and Y. Yan, "RDD-YOLO: Road Damage Detection Algorithm Based on Improved You Only Look Once Version 8," *Applied Sciences*, vol. 14, no. 8, p. 3360, Apr. 2024.

[12] A. A. Sami, S. Sakib, K. Deb, and I. H. Sarker, "Improved YOLOv5-Based Real-Time Road Pavement Damage Detection in Road Infrastructure Management," *Algorithms*, vol. 16, no. 9, p. 452, Sep. 2023.

[13] Y. Safyari, M. Mahdianpari, and H. Shiri, "A Review of Vision-Based Pothole Detection Methods Using Computer Vision and Machine Learning," *Sensors (Basel, Switzerland)*, vol. 24, no. 17, p. 5652, Aug. 2024.

[14] M. L. Pawelczyk and M. Wojtyra, "Real World Object Detection Dataset for Quadcopter Unmanned Aerial Vehicle Detection," *IEEE Access*, vol. 8, pp. 174 394–174 409, Jan. 2020.

[15] Q. Tao, H. Yang, and J. Cai, "Exploiting Web Images for Weakly Supervised Object Detection," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1135–1146, May 2019.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[17] M. Hussain, "YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection," *Machines*, vol. 11, no. 7, p. 677, Jun. 2023.

[18] R. Khanam, M. Hussain, R. Hill, and P. Allen, "A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications," *IEEE Access*, vol. 12, pp. 94 250–94 295, Jan. 2024.

[19] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, Feb. 2016.

[20] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, Mar. 2024.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.

[23] K. He, S. Ren, X. Zhang, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017.

[25] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning." *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1562–1573, Jan. 2018.

[26] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, Mar. 2016.

[27] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai, H. Machino, K. Kobayashi, K. Asada, M. Komatsu, S. Kaneko, M. Sugiyama, and R. Hamamoto, "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," *Journal of Medical Systems*, vol. 48, no. 1, Sep. 2024.

[28] R. Roslidar, M. J. Alhamdi, A. Rahman, K. Saddami, F. Arnia, M. Syukri, and K. Munadi, "Self-supervised bi-pipeline learning approach for high interpretation of breast thermal images," *IEEE Access*, vol. 12, pp. 103 433–103 449, 2024.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." institute of electrical electronics engineers, Jun. 2016, pp. 779–788.

[30] Q. Cheng, H. Wang, B. Zhu, Y. Shi, and B. Xie, "A Real-Time UAV Target Detection Algorithm Based on Edge Computing," *Drones*, vol. 7, no. 2, p. 95, Jan. 2023.

[31] Z. J. Khow, Y.-F. Tan, H. A. Karim, and H. A. A. Rashid, "Improved YOLOv8 Model for a Comprehensive Approach to Object Detection and Distance Estimation," *IEEE Access*, vol. 12, pp. 63 754–63 767, Jan. 2024.

[32] X. Wang, H. Gao, Z. Jia, and Z. Li, "BL-YOLOv8: An Improved Road Defect Detection Model Based on YOLOv8." *Sensors*, vol. 23, no. 20, p. 8361, Oct. 2023.

[33] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] H. Park and J. Paik, "Pyramid attention upsampling module for object detection," *Ieee Access*, 2022.

[35] J. Shang, J. Wang, S. Liu, C. Wang, and B. Zheng, "Small target detection algorithm for uav aerial photography based on improved yolov5s," *Electronics*, 2023.

[36] L. Han, F. Li, H. Yu, K. Xia, Q. Xin, and X. Zou, "Birpn-yolovx: A weighted bidirectional recursive feature pyramid algorithm for lung nodule detection," *Journal of X-Ray Science and Technology*, vol. 31, no. 2, pp. 301–317, 2023, pMID: 36617767. [Online]. Available: https://doi.org/10.3233/XST-221310

[37] W. Li, F. Heltha, A. Rahman, and M. I. Solihin, "Surface flaw detection of small automotive component based on improved faster r-cnn and yolo algorithms," *The International Journal of Advanced Manufacturing Technology*, 2026. [Online]. Available: https://doi.org/10.1007/s00170-026-18208-0

[38] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 993–13 000, Apr. 2020.

[39] H. Peng and S. Yu, "A Systematic IoU-Related Method: Beyond Simplified Regression for Better Localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5032–5044, Jan. 2021.

[40] T.-H. Cheung and D.-Y. Yeung, "A Survey of Automated Data Augmentation for Image Classification: Learning to Compose, Mix, and Generate." *IEEE transactions on neural networks and learning systems*, vol. 35, no. 10, pp. 13 185–13 205, Oct. 2024.

[41] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features." institute of electrical electronics engineers, Oct. 2019, pp. 6022–6031.

[42] T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, "Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions," *IEEE Access*, vol. 12, pp. 187 536–187 571, Jan. 2024.

[43] N. Dvornik, J. Mairal, and C. Schmid, "On the Importance of Visual Context for Data Augmentation in Scene Understanding." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2014–2028, Jan. 2020.

[44] R. Walambe, A. Marathe, and K. Kotecha, "Multiscale Object Detection from Drone Imagery Using Ensemble Transfer Learning," *Drones*, vol. 5, no. 3, p. 66, Jul. 2021.

Smarter Bridges: Leveraging Artificial Intelligence to Reshape University-Industry Technology Transfer

Khaouja Mohammed[1], Sanaa DFOUF[2], Kaoutar ERRAKHA[3], Hanan ELHARISSI[4], Fekkak HAMDI[5]
Doctor of Management Sciences-Faculty of Economics and Management-Settat, Hassan 1 University, Morocco[1]
Laboratory for Study and Research in Management of Organizations and Territories, ERMOT, FES[1]
Doctor of Economic and Management Sciences, Hassan 2 University, Mohammedia, Morocco[2]
Doctor of Artificial Intelligence and Computer Science-
Faculty of Sciences and Technologies of Settat, Hassan 1 University, Morocco[3]
Assistant Professor, Center for Educational Guidance and Planning (COPE),
Human Resource Management (HRM), Rabat, Morocco[4]
Teacher Researcher, LARPEG, Sultan Moulay Slimane University, Beni Mellal[5]

***Abstract*—University-industry technology transfer (UITT) is essential for converting academic research into commercial use, yet traditional strategies often fail to address the knowledge gap. Literature suggests that institutional inertia, communication barriers, and ineffective marketing strategies hinder the commercialization of technology. This study proposes a conceptual framework that incorporates AI-driven marketing to enhance knowledge dissemination, market identification, and stakeholder engagement within the technology transfer process. This systematic literature review amalgamates insights from UITT, AI marketing applications, and knowledge management systems. A qualitative analysis of peer-reviewed literature from 2017 to 2025 identifies trends, deficiencies, and emerging patterns, leading to an integrated framework that assesses technology transfer strategies and the implementation of AI marketing across diverse sectors, leveraging the Technology-Organization-Environment (TOE) model and the Unified Theory of Acceptance and Use of Technology (UTAUT). The investigation demonstrates that AI-enhanced marketing can significantly bolster UITT through five AI-enhanced marketing capabilities: precise client segmentation, predictive analytics of market trends, tailored communication, improved knowledge management, and streamlined digital outreach. This methodology fosters reciprocal knowledge exchanges, positioning AI as a facilitator between market insights and university research aims while refining technology presentations for industry stakeholders. Moreover, the study highlights critical concerns regarding data privacy, implementation expenses, technical complexities, and the necessary proficiency in AI and technology transfer.**

***Keywords*—Collaboration; university-industry; technology transfer; AI-enhanced marketing; innovation; knowledge sharing; economic growth; strategic partnerships; research initiatives; entrepreneurial mindset**

I. Introduction

The integration of artificial intelligence (AI) into marketing strategies and the transfer of technology between universities and industry represent a rapidly advancing field of research with far-reaching implications for global innovation ecosystems [1], [2], [3]. In today's academic landscape, higher education institutions face growing pressure to effectively commercialize research outcomes while fostering collaborative industry partnerships through AI-driven marketing approaches [4], [5], [6]. Traditional technology transfer models, which have long relied on licensing agreements and spin-off ventures, now confront significant challenges in the era of digital transformation. These shifts necessitate more intelligent, data-driven strategies for industry engagement [7], [8], [9].

Recent advancements in artificial intelligence (AI) have revolutionized marketing practices across diverse industries, enabling unprecedented levels of personalization, predictive analytics, and automated decision-making [10], [11], [27]. However, the application of AI-driven marketing strategies to technology transfer between academia and industry remains underexplored, creating a significant gap in both theoretical frameworks and practical implementation guidelines [13], [14]. This research void is particularly concerning given the critical role that effective marketing plays in facilitating knowledge exchange between research institutions and commercial enterprises [15], [16].

The complex nature of partnerships between universities and industries encompasses a multitude of stakeholders, a variety of knowledge types, and a spectrum of organizational cultures, all of which necessitate the establishment of intricate coordination mechanisms to ensure the efficacy of collaborative endeavors, as articulated by [17], [18], [19]. Conventional technology transfer offices often face significant challenges, including resource constraints, a lack of comprehensive market intelligence, and ineffective communication channels, which can hinder successful commercialization outcomes, as discussed by [20], [21], [22]. The integration of AI-enhanced marketing strategies offers promising potential solutions to confront these pressing challenges by facilitating the development of intelligent matching systems, delivering predictive analytics capable of forecasting the probability of partnership success, and providing automated functionalities for the management of collaborative relationships, as noted by [23], [24], [25].

Artificial intelligence applied to marketing has made significant progress, but its role in knowledge transfer between universities and industry remains largely unknown. This lack of a clear theoretical framework calls for in-depth reflection, integrating both the possibilities offered by AI and the specificities of collaborations between academia and business.

This study aims to explore several key areas of research:

- How can AI enhance the effectiveness of technology transfer to industrial players?
- What algorithmic mechanisms can better connect academic innovations to the concrete needs of businesses?
- How is AI changing its functioning and strategic value?
- How can intelligent marketing mitigate cultural and informational differences between researchers and industrialists?
- What contextual elements favor the adoption of AI in research valorization?
- How can AI, marketing, and technology transfer be articulated within a coherent conceptual framework?
- What indicators can be used to assess the real effectiveness of these approaches on university-industry collaborations?

II. Theoretical Framework

A. University-Industry Technology Transfer Foundations

University-industry technology transfer encompasses a comprehensive array of activities aimed at facilitating the transference of knowledge, technologies, and innovations from academic institutions to commercial applications [8], [9], [30]. Conventional methodologies have predominantly focused on formal mechanisms, such as patent licensing, the establishment of spin-off enterprises, and the formulation of collaborative research agreements, as illustrated in Fig. 1 [6], [7], [4]. Nevertheless, contemporary scholarship suggests that effective technology transfer requires more nuanced methodologies that address institutional, interpersonal, and cultural barriers that often hinder commercialization endeavors [8], [22], [15].

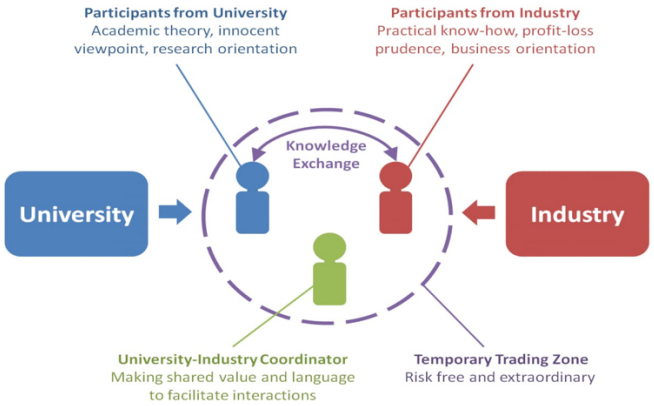


Fig. 1. University-industry collaboration for knowledge exchange.

The academic discourse delineates three principal categories of inhibitors that obstruct the successful commercialization of research-based inventions: institutional factors pertinent to organizational structures and policies, interpersonal dynamics that pertain to the relationship interactions between academic and industry stakeholders, and cultural disparities that shape collaboration methodologies [8], [20], [26]. Recent investigations underscore that relational inhibitors constitute the most prevalent challenges, while institutional and cultural factors frequently exacerbate these relational impediments [8], [7], [15]. These findings underscore the paramount importance of devising enhanced communication and engagement strategies that can overcome traditional barriers to the effectiveness of technology transfer.

Empirical investigations across diverse international contexts exhibit substantial discrepancies in technology transfer performance across various institutional and regional frameworks [6], [30], [22]. Collaborative networks between Chinese universities and industries manifest small-world phenomena characterized by distinct centrality patterns that influence innovation outcomes [30], [15]. Emerging economies in Africa face specific challenges stemming from resource limitations and inadequate technological infrastructure, which significantly impact the effectiveness of technology transfer [22], [38], [40]. These contextual variations underscore the need for adaptable frameworks that can accommodate diverse institutional environments while preserving the core principles of effectiveness.

B. Artificial Intelligence in Marketing Strategies

The incorporation of artificial intelligence within marketing methodologies signifies a profound transition from conventional techniques toward data-centric, automated, and individualized engagement frameworks [3], [2], [27]. Modern AI marketing implementations include machine learning algorithms for effective customer segmentation, natural language processing for content enhancement, predictive analytics for forecasting consumer behavior, and chatbot technologies for automated customer support [10], [36], [14]. These technological advancements collectively empower organizations to execute highly targeted, efficient, and scalable marketing initiatives that respond to the unique preferences of customers and the evolving market conditions. Table I delineates a comparison between conventional marketing approaches and those augmented by artificial intelligence [2], [1], [23]. Empirical research indicates that artificial intelligence substantially augments marketing efficacy through enhanced automation, personalization capabilities, and precision in forecasting [3], [2], [14]. Organizations that adopt AI-enhanced marketing methodologies report significant advancements in operational efficiency, customer engagement metrics, and competitive standing within their respective markets [1], [27]. Nevertheless, the existing literature also highlights significant challenges, including elevated implementation expenditures, ethical dilemmas related to data privacy, and the need for specialized technical proficiency, which may hinder widespread adoption [10], [36], [13].

Comparative analysis between developed and developing nations highlight notable discrepancies in AI marketing adoption trends, influenced by factors such as infrastructure preparedness, human resource competencies, and supportive policy frameworks, as illustrated in Fig. 2 [13], [46], [41]. Typically, developed countries exploit AI for advanced applications such as content personalization, predictive analytics, and marketing automation, whereas developing nations frequently encounter barriers associated with insufficient digital infrastructure and low technical literacy levels [13], [39], [45]. These

TABLE I. TRADITIONAL VS. AI-ENHANCED MARKETING

Aspect / Dimension	Traditional Approach	AI-Enhanced Approach	Benefits of AI
Market ID	Manual research, surveys, and conventional market studies	Predictive analysis, data mining, automated identification	Increased accuracy, emerging market detection, early identification of new opportunities
Lead Generation	Events, networking, mass marketing	Smart targeting, predictive scoring, process automation	Improved lead quality, increased efficiency
Content Personalization	Broad segments, generic messages	Hyper-personalization, dynamic and adaptive content	Higher engagement, better conversion rates
Communication Timing	Fixed schedules, sequential approach	Optimal timing based on behavioral data	Maximized receptivity, reduced intrusion
Success Measurement	Static indicators (KPIs), periodic reports	Real-time analytics, outcome prediction	Immediate adjustments, continuous campaign optimization
Resource Allocation	Fixed budgets, uniform distribution	Dynamic allocation based on performance	Maximized ROI, resource efficiency
Stakeholder Targeting	Static lists, demographic segmentation	Behavioral profiles, predictive targeting	Increased relevance, reduced acquisition costs
Technology Positioning	Standardized messages, one-size-fits-all	Adaptive positioning based on audience	Improved understanding, facilitated adoption
Follow-up Processes	Manual tracking, scheduled follow-ups	Automated nurturing, trigger-based engagement	Consistent engagement, reduced drop-offs
Decision Support	Experience-based, intuitive decisions	Data-driven recommendations, scenario modeling	Reduced bias, improved outcomes

findings underscore the importance of contextual elements in developing effective AI marketing implementation strategies.

The bibliometric assessment of AI marketing literature between 2021 and 2025 reveals an escalating scholarly interest in this research domain. The review was conducted following the PRISMA methodology to ensure a transparent, systematic, and reproducible article selection process. The analysis identified contributions from 82 nations and 1,292 authors across 281 sources, underscoring the global relevance and growing academic importance of AI-driven marketing research. Fig. 3 illustrates technology transfer approaches [1], [46], [41].

The literature selection process was based on predefined inclusion and exclusion criteria, focusing on peer-reviewed studies published in English between 2021 and 2025 and directly related to artificial intelligence applications in marketing. Core cognitive frameworks encompass consumer behavior analysis, brand management, marketing strategy formulation, and digital transformation initiatives, which collectively influence contemporary AI marketing practices [1], [38].

C. Digital Transformation in Higher Education

Digital transformation within the realm of higher education entails a holistic reorganization that integrates digital technologies across all facets of institutional functioning, significantly reshaping how universities provide education, perform research and interact with external constituents [16], [25]. This transformative endeavor requires strategic coherence between technological capabilities and institutional objectives, necessitating substantial investments in physical infrastructure, human capital, and procedural innovation [19], [31]. Institutions engaging in digital transformation are tasked with reconciling traditional academic principles with contemporary technological demands, all while preserving educational standards and research integrity [5], [47] . The various maturity levels of the Technologie Knowledge Transfer University Industrie (TKTUI) are delineated in the accompanying Table II.

Current digital transformation efforts in academic institutions emphasize various dimensions, including online learning environments, data analytics systems, cloud computing infrastructure, and AI-enhanced administrative frameworks [5]. The goal of these technological advancements is to boost operational efficiency, enrich student experiences, and fortify research capabilities while fostering enhanced collaboration with external entities [25], [31]. However, the realization of successful digital transformation transcends mere technological integration; it necessitates a cultural shift, unwavering leadership dedication, and a strategic vision that permeates all levels of the organization [16], [19], [47]. The emergence of the COVID-19 pandemic expedited the adoption of digital transformation across higher education institutions, compelling the swift deployment of virtual collaboration tools and online service delivery systems [48], [42]. This rapid transition uncovered both opportunities and challenges linked to digital transformation, such as enhanced accessibility and flexibility juxtaposed with concerns regarding barriers to technology adoption and issues of digital equity [32], [48], [35]. These occurrences provide valuable insights into the factors that influence successful digital transformation within the university context. The subsequent Table III enumerates the components associated with digital transformation in higher education institutions.

D. Knowledge Management and Transfer Mechanisms

Knowledge management within the domain of technology transfer encompasses intricate methodologies for the acquisition, structuring, dissemination, and utilization of both explicit and tacit knowledge across organizational frontiers [7], [17], [20] . Conventional technology transfer paradigms tend to prioritize formal intellectual property frameworks such as patents and licensing contracts; however, successful knowledge transfer necessitates holistic strategies that adequately address the considerable tacit knowledge elements intrinsic to academic technologies [7], [15], [20]. The tacit aspect can be particularly pronounced due to the nascent character of academic technologies and the specialized knowledge requisite for their practical application [7], [18], [24]. Contemporary studies underscore the essential function of human and institutional networks in facilitating the transmission of tacit knowledge between academic institutions and commercial entities [7], [15], [17]. Collaborative frameworks, consulting partnerships, sponsored research initiatives, proof-of-concept facilities, and university spin-offs represent vital mechanisms that foster direct interactions conducive to tacit knowledge

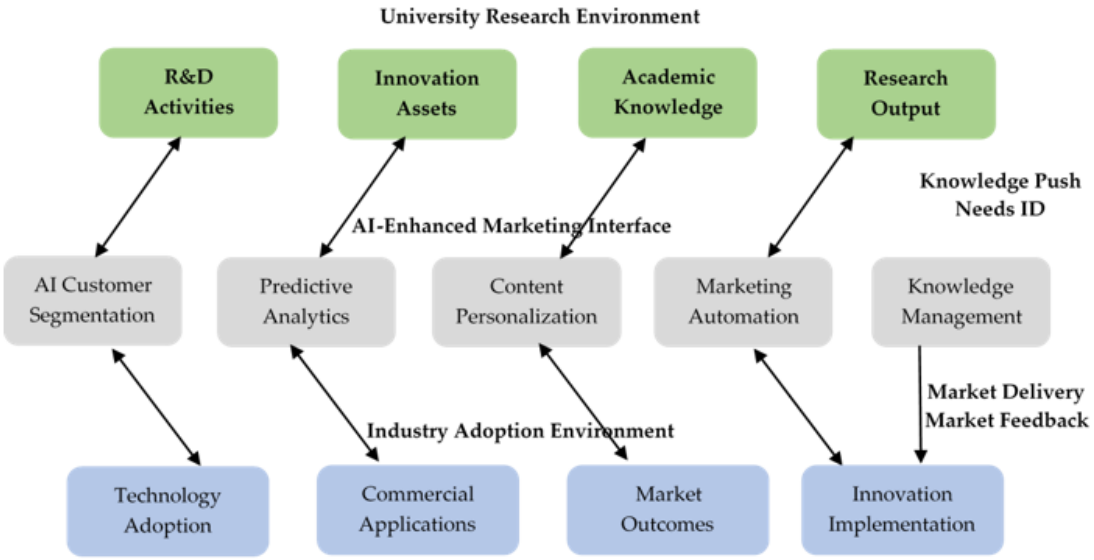


Fig. 2. Concept of AI-enhanced marketing for University-Industry Technology Knowledge Transfer.

TABLE II. DIGITAL TRANSFORMATION MATURITY LEVELS

Maturity Level	Characteristics	Technology Transfer Implications
Level 1: Basic	Limited digital infrastructure, manual processes, basic IT systems	Basic communication tools, limited market reach, manual relationship management
Level 2: Developing	Some digital tools implemented, inconsistent processes, emerging digital awareness	Enhanced communication, simple database management, emerging online presence
Level 3: Defined	Standardized digital processes, integrated systems, and clear digital strategy	Integrated CRM systems, digital marketing capabilities, structured knowledge management
Level 4: Managed	Data-driven decision-making, advanced analytics, performance monitoring	Predictive analytics for partnerships, automated matching systems, performance dashboards
Level 5: Optimized	AI-enhanced processes, continuous optimization, innovation leadership	AI-driven partner identification, intelligent knowledge transfer, adaptive learning systems

transfer [7], [19], [20]. These mechanisms enhance formal licensing agreements by offering avenues for prolonged collaboration and knowledge advancement that bolster successful commercialization endeavors [17], [18], [24]. The advent of AI-augmented knowledge management systems presents novel opportunities for enhancing knowledge discovery, alignment, and transfer processes [17], [25], [41]. Sophisticated algorithms are capable of scrutinizing substantial quantities of research outputs to uncover commercialization prospects, align university technologies with industry demands, and forecast probabilities of successful collaborations [18], [19]. Nevertheless, the incorporation of AI technologies into knowledge management frameworks necessitates meticulous attention to data integrity, algorithmic transparency, and human oversight to ensure efficacious and ethical application [39], [45]. The classification of mechanisms facilitating knowledge transfer, incorporating the influence of artificial intelligence, is illustrated in the subsequent Table IV.

The mechanism categories described in Table IV emerged from an integrative thematic synthesis of the research literature on university-industry technology transfer and knowledge management. Instead of using one of the existing taxonomies, the classification system draws together the transfer and collaboration practices documented in the studies reviewed. The mechanism categorization captures the dominant logic and institutional role. This was accomplished by identifying and grouping transfer practices, resulting in five mechanism categories: formal transfer of intellectual property, spin-offs, research collaboration, human capital transfer, and informal knowledge transfer. The categorization drew on the persistent themes in the literature on technology transfer, knowledge management, and innovation ecosystems (e.g., Lee, 2020; Yazdani, 2024; Sviatetskyi, 2020; Eidlisz et al., 2024). The AI enhancement dimension was added by recognizing the role of AI in enhancing each transfer mechanism, conceptualized as automation, prediction, intelligent matching, and knowledge discovery.

III. MATERIALS AND METHODS

This research utilized an integrative literature review methodology adhering to established protocols for systematic knowledge synthesis and the formulation of conceptual models [21], [33], [29]. The integrative review framework was chosen for its ability to amalgamate various research methodologies, theoretical frameworks, and empirical evidence into a holistic comprehension of the phenomenon being examined [34], [28]. This methodology facilitates the discernment of patterns, discrepancies, and interconnections across multiple domains while fostering the creation of innovative theoretical frameworks that synthesize existing knowledge [37]. The systematic search approach employed the Scopus database as the principal resource owing to its extensive coverage of high-quality,

TABLE III. DIGITAL TRANSFORMATION ELEMENTS IN HIGHER EDUCATION

Element	Description	Role in Technology Transfer
Digital Infrastructure	Hardware, software, and network systems that support digital operations	Enables efficient information exchange and communication
Data Analytics Capabilities	Tools and expertise for analyzing and interpreting large datasets	Provides insights into research trends and market opportunities
Virtual Collaboration Platforms	Platforms that enable remote collaboration and knowledge sharing	Facilitates engagement between university and industry stakeholders
Cloud Computing Technologies	Scalable computing resources accessible via the internet	Offers scalable resources for data-intensive research and applications
AI-Enhanced Administrative Systems	Automated systems for administrative tasks and decision support	Streamlines administrative processes in technology transfer
Digital Content Management	Systems for creating, storing, and distributing digital content	Supports effective knowledge dissemination and documentation
Cybersecurity Framework	Protocols and systems for protecting digital assets and privacy	Protects intellectual property and sensitive information

TABLE IV. KNOWLEDGE TRANSFER MECHANISMS TAXONOMY

Mechanism Type	Specific Mechanisms	AI Enhancement Potential
Formal IP Transfer	Patent licensing, technology licensing, copyright transfer	Automated IP matching, valuation algorithms, smart contract management
Spin-off Creation	Academic entrepreneurship, university spin-offs, startup incubation	Market opportunity identification, funding source matching, risk assessment
Collaborative Research	Joint research projects, industry-sponsored research, research consortia	Partner matching algorithms, research trend analysis, collaboration optimization
Human Capital Transfer	Student internships, faculty consulting, staff exchange programs	Skill matching, expert identification, knowledge gap analysis
Informal Knowledge Exchange	Conferences, workshops, networking events, publications	Intelligent networking, content recommendation, knowledge discovery

peer-reviewed literature pertinent to technology management, marketing, and artificial intelligence scholarship. The search methodology employed various keyword combinations, including "artificial intelligence marketing", "university-industry technology transfer", "digital marketing strategies", "knowledge transfer mechanisms", and "innovation ecosystems", to ensure thorough coverage of the pertinent literature. Boolean operators and controlled vocabulary terms were utilized to enhance search accuracy while ensuring sensitivity to emerging research domains (see Table V).

The study selection process consisted of multiple screening phases, including title and abstract examination, full-text analysis, and quality evaluation based on predefined inclusion and exclusion criteria. The inclusion criteria encompassed peer-reviewed articles published in English, as well as empirical investigations and theoretical contributions related to AI, marketing, or technology transfer, and publications from 2017 to 2025 to capture contemporary advancements. Exclusion criteria ruled out conference proceedings, books, editorial notes, and studies not directly pertinent to the research objectives.

Data extraction and synthesis methodologies adhered to systematic protocols for identifying key themes, concepts, methodological approaches, and theoretical contributions across the selected studies. Thematic analysis techniques were employed to identify patterns and relationships within and between studies, facilitating the development of comprehensive conceptual frameworks that integrate diverse perspectives. Quality assessment procedures appraised methodological rigor, theoretical contributions, and practical relevance to ensure the reliability and validity of the synthesized outcomes.

The results section has been organized as per the sequential workflow illustrated in Table V. More specifically, findings are organized in relation to the various methodological phases of the study, namely study selection, bibliometric classification, thematic synthesis, analytical mapping, and the construction of the conceptual model. This was done to achieve a consistent methodology, enhance readability, and offer a clear link between the review process and the resultant conceptual framework.

IV. RESULTS AND DISCUSSION

Research articles were considered eligible for inclusion if they met the following criteria: 1) published in peer-reviewed scholarly journals or reputable conference proceedings, 2) focused on AI-enhanced marketing, technology transfer between universities and industry, or the intersection of these domains, 3) published between 2017 and 2025, and 4) written in English. The exclusion criteria included: 1) non-academic publications, 2) research lacking methodological rigor, 3) duplicate publications, and 4) studies beyond the defined area, as shown in PRISMA Flow Diagram (see Fig. 3).

A. Study Selection and Temporal Distribution

This subsection addresses the methodological foundation underlying RQ1 through RQ7 by establishing the evidentiary basis from which answers to all seven research questions are

TABLE V. RESEARCH METHODOLOGY

Phase	Description	Key Activities	Expected Outcomes
Phase 1: Systematic Search	A comprehensive search of Scopus, Web of Science, IEEE Xplore, Google Scholar, SpringerLink, Emerald Insight, and peer-reviewed databases using systematic keywords	Database search; Keyword optimization; Search strategy development	Comprehensive literature corpus relevant to AI-enhanced marketing for technology transfer
Phase 2: Study Selection	Application of inclusion/exclusion criteria, screening abstracts, and full texts	Duplicate removal; Title/abstract screening; Full-text evaluation	High-quality studies meeting inclusion criteria for detailed analysis
Phase 3: Data Extraction	Extraction of key data elements, themes, and concepts from selected studies	Data coding; Theme identification; Concept mapping	Structured data matrix with key themes, concepts, and relationships
Phase 4: Quality Assessment	Assessment of study quality using established criteria and frameworks	Bias assessment; Methodological evaluation; Evidence grading	Quality-assessed evidence base for reliable synthesis
Phase 5: Data Synthesis	Thematic analysis and synthesis of findings using an integrative approach	Pattern identification; Cross-case analysis; Framework development	An integrated understanding of the current state and research trends
Phase 6: Model Development	Development of a conceptual model based on synthesized evidence	Model construction; Component integration; Relationship mapping	Novel conceptual model integrating AI, marketing, and technology transfer

derived. The temporal trend of publication growth corroborates the growing scholarly relevance of the questions posed.

The systematic selection process was designed to ensure transparency and reproducibility in the identification and screening of relevant studies. The search encompassed publications addressing AI-mediated university-industry marketing strategies over the preceding decade. A total of 48 papers met the final inclusion criteria. The temporal distribution of included studies revealed an increase in publication output beginning in 2022, with the highest annual output recorded in 2024 (18 publications). This trend corresponds to the growing body of research at the intersection of university-industry technology transfer (UITT) and artificial intelligence (AI) [43].

B. Bibliometric Classification of Included Studies

This subsection primarily addresses RQ3 (How is AI changing its functioning and strategic value?) by quantifying the thematic distribution of the literature and documenting the dominant trajectory of AI adoption in marketing. It also provides foundational evidence for RQ1 by establishing the volume and focus of AI-enhanced marketing research. To provide the highest level of transparency and reproducibility, a systematic manual coding approach was applied to the bibliometric classification. The 48 selected studies were independently analyzed and classified based on two different dimensions: 1) the main research theme, and 2) the research methodology. The thematic classification was based on a constructivist approach and was an outcome of the inductive analysis of the main objectives of the reviewed articles and the main research questions, key concepts, and main contributions. The first phase of the coding process was open and resulted in a number of key concepts related to the implications of artificial intelligence, the university-industry technology transfer, the process of digital transformation, the management of knowledge and the innovation management ecosystem. These related concepts were combined to form larger categories, which represented the main focus of each of the studies.

In the classification of the research methodology, studies were differentiated based on the main research method that formed the basis of the studies, which included the case study method, the formulation of a conceptual/theoretical paper, empirical quantitative research, qualitative research, bibliometric research, and modeling and predictive research. The classification was based on the explanation of the research methodology that was included in each study.

The manual classification relied on the fact that a number of the studies addressed themes that were multiple and interrelated, and that a context was necessary in order to explain the interrelated aspects of each of the studies. To strengthen the consistency of the classification, a coding scheme that was derived from the objectives of the review and the research questions, that is, RQ1-RQ7, was used to provide guidance as to the classification. For studies that focused on numerous themes, the main analytical focus was the primary theme to which the study was assigned, and the remaining themes were recorded in the process of thematic synthesis.

The 48 included studies were classified according to their primary thematic focus and methodological approach. AI-driven marketing strategies constituted the largest thematic category (47.8%), followed by studies addressing institutional barriers to information and communication technology adoption (30.4%). Studies focused on methodological advancements represented a smaller proportion (6.5%), indicating limited coverage of standardized analytical frameworks in the existing literature. Within the AI-focused subset, studies were further classified by technology application type: personalization (40.9%), automation (27.3%), and predictive analytics (18.2%). Case study designs were the most prevalent methodology among the included papers (58.7%), particularly in marketing-related studies, while mathematical modeling approaches were underrepresented.

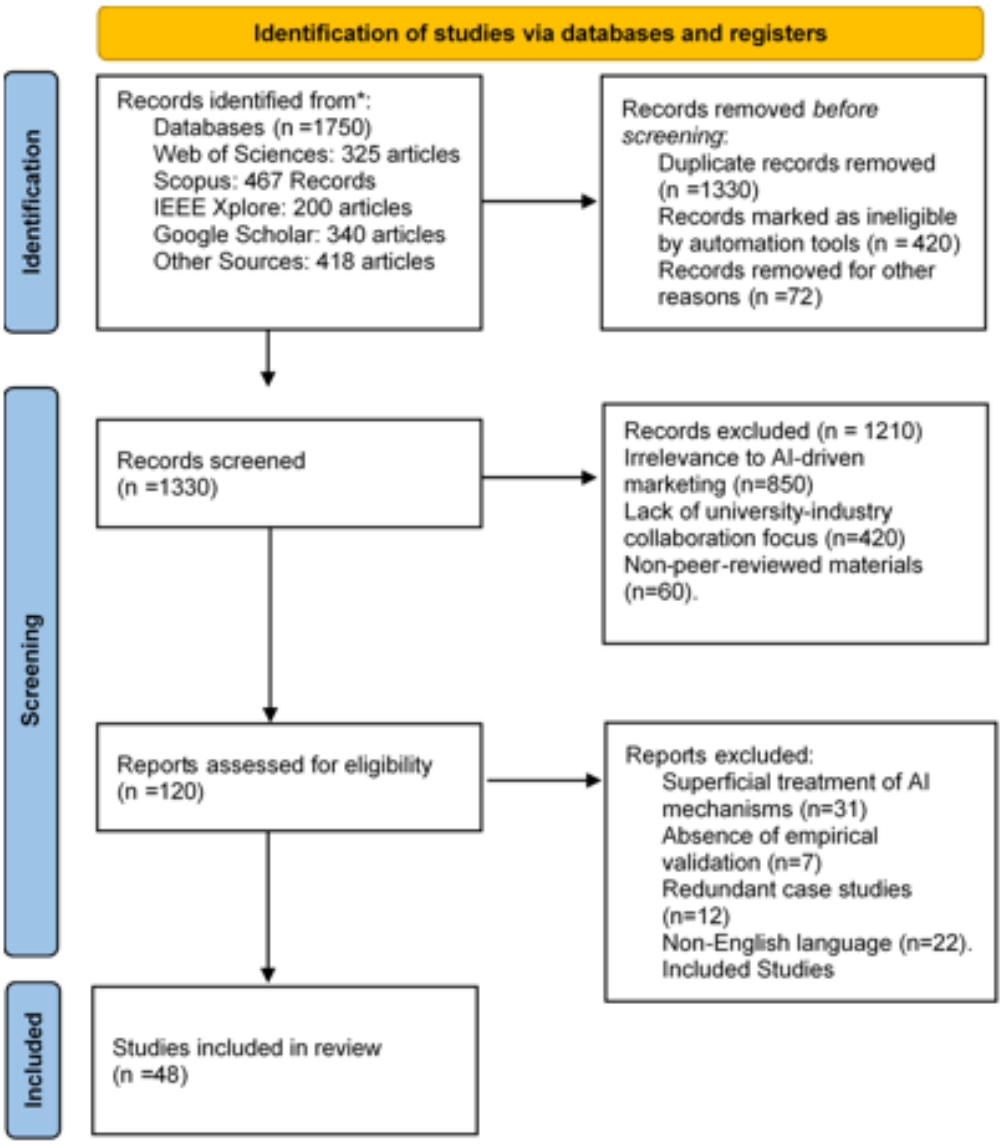


Fig. 3. PRISMA flow diagram

While examining the literature, the specified capabilities of segmentation, predictive analytics, personalized messaging, intelligent knowledge management, and automated digital engagement, were frequently observed within the context of the thematic synthesis process. The listed capabilities were also the primary functional contributions that the marketed enhancements attributed to the technology transfer activities of higher education institutes. These capabilities of AI are distinct from the marketing mechanisms of knowledge transfer. The former articulates the operational mechanisms that artificial intelligence augments within marketing and collaboration, while the latter describes the pathways of institutional and organizational knowledge transfer mechanisms, such as patent licenses, spin-offs, partnerships in research, and consultancy.

C. Thematic Synthesis Procedure

This subsection provides the analytical bridge connecting the raw literature findings to the answers for all seven research questions. The identification of seven major thematic categories directly maps onto the seven research questions, as each theme constitutes the evidential basis for one or more framework components developed in Section V.

To bridge the gap between descriptive bibliometric results and the development of a conceptual framework, a thematic synthesis was conducted following established qualitative synthesis procedures. This approach involved three stages: 1) line-by-line coding of key findings from each of the 48 included papers; 2) the organization of codes into descriptive themes; and 3) the generation of analytical themes informing

framework component development.

A total of 131 theme assignments were recorded across 48 papers, yielding an average of 2.73 themes per paper. This multi-dimensional coding reflected the interconnected nature of the concepts addressed in the literature and informed the identification of seven major thematic categories.

D. Analytical Principles for Theme-to-Component Mapping

This subsection supports RQ6 (How can AI, marketing, and technology transfer be articulated within a coherent conceptual framework?) by establishing the transparent methodological criteria used to translate thematic evidence into discrete, justifiable framework components. Three analytical principles guided the mapping of synthesized themes to the proposed framework components:

- Principle 1 – Frequency-Based Prominence: Themes with literature coverage exceeding 30% were designated as candidates for dedicated framework components.
- Principle 2 – Functional Differentiation: Themes with moderate coverage were retained as framework components when they served functionally distinct roles not addressed by higher-frequency themes.
- Principle 3 – Integration of Cross-Cutting Insights: Findings from multi-dimensional papers (i.e., those coded under multiple themes) were used to identify interdependencies and inform the relational architecture of the framework.

V. Conceptual Model Development

The proposed AI-enhanced marketing Framework for University-Industry Technology Knowledge Transfer integrates eight interconnected components that collectively optimize collaboration effectiveness and commercialization outcomes.

A. Framework Component Derivation

This subsection directly addresses RQ6 in its entirety, and provides evidence for RQ1 and RQ2 through the derivation of the AI Technologies (Component 1) and Marketing Strategies (Component 2) components. The bilateral treatment of University Factors (Component 3) and Industry Factors (Component 4) addresses RQ4 by explaining the structural sources of cultural and informational differences between academic and industrial stakeholders (see Table VI).

The application of the above principles to the thematic synthesis results yielded eight framework components. The derivation procedure for each component is described below, including the supporting literature base, thematic evidence, and justification for inclusion.

1) Component 1 - AI Technologies: This component was derived from 23 papers (47.8% of the corpus). The thematic coding identified AI-enhanced marketing as the most frequently addressed theme. Sub-theme analysis revealed three distinct capability clusters: personalization (40.9%; [1], [27], [23]), automation (27.3%; e.g., [2], [14]), and predictive analytics (18.2%; e.g., [10], [36]). The high frequency of this theme, combined with the differentiated capability clusters, justified its inclusion as a foundational component. The sub-element structure (machine learning, natural language processing, predictive analytics, chatbots, automation) directly mirrors the identified clusters. Technical capabilities were distinguished from strategic applications to warrant separate treatment from Component 2. This component constitutes the primary answer to RQ1 by identifying the specific AI mechanisms (machine learning, NLP, predictive analytics, chatbots, and automation) through which technology transfer effectiveness is enhanced.

2) Component 2 - Marketing Strategies: This component was informed by 18 papers (37.5%). The thematic evidence encompassed the strategic deployment of AI in university-industry contexts, digital transformation in marketing, stakeholder engagement optimization, and content personalization for industry partners. While thematically adjacent to Component 1, these papers addressed strategic deployment rather than technical capabilities, and the specialized marketing approaches required for the UITT context differed from conventional B2B models (e.g., [12], [16], [3], [13], [24]). This component answers RQ2 by specifying the algorithmic and strategic mechanisms — including intelligent matching, predictive partner scoring, and behaviorally-driven content personalization — that connect academic innovations to concrete industry needs.

3) Component 3 - University Factors: This component drew on 19 papers (39.6%). The thematic coding aligned with the UITT Barriers theme, where institutional factors accounted for 30.4% of the literature. Three sub-themes emerged: organizational structures, research capabilities, and intellectual property management. The identification of institutional barriers as the second most frequent theme, combined with the decomposition into actor-specific factors, justified dedicated component status (e.g., [8], [9], [4], [6], [7], [20]). This component contributes to answering RQ4 by identifying the institutional sources of informational gaps and cultural barriers on the academic side, including misaligned incentive structures, limited industry-facing communication, and inadequate IP commercialization capacity.

4) Component 4 - Industry Factors: This component was supported by 15 papers (31.2%). The thematic evidence addressed industry readiness, absorptive capacity, technology readiness, collaborative willingness, and trust mechanisms. Inclusion was justified on the basis of bilateral representation, balancing the university-side factors in Component 3. The synthesis revealed absorptive capacity, technology readiness, and collaborative willingness as distinct yet interrelated dimensions (e.g., [17], [30], [36], [22], [18]). In conjunction with Component 3, this component completes the answer to RQ4 by characterizing the industry-side cultural and informational barriers — particularly low absorptive capacity and insufficient trust in academic partners — that AI-enhanced marketing strategies must actively mitigate.

5) Component 5 - Knowledge Transfer Mechanisms: This component was derived from 17 papers (35.4%), corresponding to the Knowledge Management theme. The coding distinguished between tacit and explicit knowledge transfer modes, as well as formal mechanisms (patents, licensing, spin-offs) and informal mechanisms (consulting, collaborative research,

TABLE VI. DETAILED MAPPING TABLE

Component	Sub-elements	Primary Themes	Supporting Evidence	Key Citations	Analytical Justification
1. AI Technologies	Machine Learning, NLP, Predictive Analytics, Chatbots, Automation	AI Marketing; Personalization; Automation	Nuță (2024), Gabelaia (2024), Potwora et al. (2024)	23 papers (47.8%)	Nearly half of studies focus on AI, especially personalization and automation.
2. Marketing Strategies	Digital Marketing, Content Personalization, Segmentation, Social Media	AI marketing; Digital communication	Nektarios (2022), Ahmad et al. (2024)	18 papers (37.5%)	Shift to AI-driven strategies.
3. University Factors	Research, TTOs, Expertise, IP	Institutional barriers; Capacity	Battaglia et al. (2024); Blankesteijn et al. (2021)	19 papers (39.6%)	Internal factors affect knowledge transfer success.
4. Industry Factors	Market needs, Capacity, Readiness	Market dynamics; Collaboration	Yazdani (2024); Jiang et al. (2022)	15 papers (31.2%)	External factors complement university capabilities.
5. Knowledge Transfer Mechanisms	Licensing, Spin-offs, Consulting	Knowledge transfer	Lee (2020); Yazdani (2024)	17 papers (35.4%)	Bridge between university and industry.
6. Digital Transformation	Infrastructure, Data, Platforms	Digital transformation	Nektarios (2022); Ahmad et al. (2024)	16 papers (33.3%)	Enables AI and marketing.
7. Innovation Ecosystem	Networks, Partnerships, Policy	Ecosystems	Fadeyi et al. (2019)	12 papers (25.0%)	Impacts scalability.
8. Performance Outcomes	Commercialization, Innovation	Metrics	Singh et al. (2023)	11 papers (22.9%)	Supports evaluation.

networking). This component emerged as a conceptual bridge between university and industry actors. The comprehensive taxonomy of transfer mechanisms identified across 17 papers provided both theoretical and empirical grounding (e.g., [7], [20], [17], [15], [19], [24]). This component addresses RQ1 and RQ2 by identifying the specific pathways through which AI-augmented tools — such as partner-matching algorithms and AI-assisted IP valuation — optimize the operationalization of knowledge transfer.

6) Component 6 - Digital Transformation elements: This component was informed by 16 papers (33.3%). Digital transformation appeared as a cross-cutting theme encompassing infrastructure requirements, data analytics capabilities, cloud computing, and virtual collaboration platforms. This component was distinguished from AI Technologies (Component 1) as it represents broader digital readiness rather than specific AI capabilities. The inclusion of maturity level progression provided a developmental dimension (e.g., [12], [5], [31], [47], [16]). This component addresses RQ5 by identifying the enabling contextual conditions — digital infrastructure maturity levels, cloud adoption, and platform readiness — that govern the feasibility and effectiveness of AI adoption in research valorization activities.

7) Component 7 - Innovation Ecosystem: This component drew on 12 papers (25.0%). The thematic evidence addressed external contextual factors, including policy frameworks, funding mechanisms, regional contexts, and collaborative network structures. Inclusion was justified by the Frequency-Based Prominence principle (meeting the 25% threshold) and by the finding that comparative analyses across studies revealed ecosystem-level moderators that significantly affect technology transfer effectiveness, particularly in emerging economy contexts (e.g., [22], [6], [40], [38], [44]). This component further addresses RQ5 by documenting the broader ecosystem-level elements — government support, incubation infrastructure, and international networks — that determine the contextual readiness for AI adoption in technology transfer, particularly in emerging economy contexts.

8) Component 8 - Performance Outcomes: This component was supported by 11 papers (22.9%), with performance measurement addressed explicitly in these papers and implicitly throughout the broader corpus. The thematic evidence covered technology commercialization metrics, knowledge transfer efficiency, innovation performance indicators, and competitive advantage evaluation. Inclusion was justified by the component's feedback function, which enables continuous improvement and empirical validation of framework effectiveness (e.g., [21], [41], [30], [18], [19], [38]). This component constitutes the direct answer to RQ7 by specifying the indicators — including commercialization rates, collaboration success scores, knowledge transfer efficiency indices, and innovation impact metrics — that can be used to assess the real effectiveness of AI-enhanced marketing approaches on university-industry collaborations.

B. Framework Validation Criteria

This subsection supports RQ6 by providing a systematic validation of the proposed framework against four criteria that confirm its suitability as a coherent theoretical articulation of the relationships among AI, marketing, and technology

transfer. The resulting eight-component framework was assessed against four validation criteria to confirm its systematic derivation from the literature:

- Empirical foundation: Each component was supported by 11 to 23 papers (22.9%–47.8% coverage).
- Functional distinctiveness: Each component occupies a different structural role within the framework (enablers, actors, processes, moderators, or outcomes).
- Research integration: Each component synthesizes previously fragmented research streams into a coherent dimension.
- Research generativity: The framework structure provides testable propositions and clear boundaries for future empirical investigation. This thematic synthesis procedure, combined with the three analytical principles and the component derivation process, establishes an explicit evidence trail from the 48 included papers to the eight-component conceptual framework for AI-enhanced university-industry technology transfer.

C. Core Framework Components

This subsection integrates answers to RQ1, RQ2, RQ3, and RQ4. It describes how AI Technologies (RQ1, RQ3) and Marketing Strategies (RQ2, RQ4) interact with university and industry factors to generate the core capabilities of the proposed framework. This comprehensive framework integrates findings from the literature review to address recognized deficiencies in both theoretical understanding and practical implementation strategies. The model (Fig. 4) delineates Knowledge Transfer Mechanisms as the pivotal core around which all supplementary elements engage, underscoring the essential significance of efficient transfer processes in realizing fruitful collaborations between academia and industry. AI Technologies acts as the fundamental catalyst for intelligent automation and decision-making mechanisms within the technology transfer ecosystem [3], [2], [1]. This aspect includes machine learning models for recognizing patterns and making predictions, natural language processing for analyzing and generating content, predictive analytics for forecasting the success of collaborations, chatbot technologies for automating stakeholder interactions, and systems for optimizing processes [27], [36], [14]. The amalgamation of these AI technologies fosters opportunities for heightened efficiency, diminished manual labor, and enhanced precision in decision-making across the entirety of university-industry collaborations [10], [23].

In direct response to RQ1, these capabilities demonstrate how AI can enhance technology transfer effectiveness by automating partner discovery, accelerating IP matching, and enabling data-driven decision-making at every stage of the collaboration lifecycle.

Marketing Strategies signify the strategic communication and engagement methodologies that universities utilize to draw, engage, and sustain relationships with industry collaborators [12], [16], [25] . This dimension encompasses digital marketing initiatives designed explicitly for industry stakeholders, content personalization based on partner inclinations and requirements; customer segmentation approaches that pinpoint optimal collaboration prospects, social media platforms for fostering relationships, and data-informed marketing techniques that enhance resource distribution and campaign efficacy [3], [1], [13] . These strategies ought to be tailored to the distinctive attributes of technology transfer settings while capitalizing on AI functionalities for improved accuracy and influence [2], [14], [31]. In response to RQ2 and RQ3, these strategies illustrate how AI is transforming the strategic value of marketing from a static broadcast mechanism into a dynamic, predictive, and relationship-intelligent function. University Factors include the internal competencies, resources, and traits that affect the success of technology transfer initiatives [6], [5]. Research capabilities dictate the caliber and commercial viability of university technologies, whereas Technology Transfer Offices furnish the institutional framework for overseeing collaboration processes [20]. Academic expertise embodies the human resources requisite for efficient knowledge transfer, intellectual property portfolios lay the groundwork for commercialization endeavors, and an innovation-oriented culture impacts organizational preparedness for external collaborations [7], [15], [19]. These elements must be strategically cultivated and aligned to bolster AI-augmented marketing initiatives [17], [18]. This component contributes to answering RQ4 by identifying the academic institutional factors — particularly cultural resistance to commercialization and weak industry-facing communication infrastructure — that intelligent marketing strategies must specifically address.

Industry Factors exemplify the external market dynamics, organizational attributes, and collaboration readiness elements that affect the success of partnerships. Market demands to propel the necessity for university technologies and ascertain commercialization feasibility, while absorptive capacity denotes the capability of industry partners to comprehend, adapt, and execute university innovations [17], [19], [20]. Technology readiness influences the timing and practicality of collaborative initiatives, collaborative willingness affects the formation and sustainability of partnerships, and resource availability dictates the extent and scale of potential collaborative activities [15], [47]. A comprehensive understanding and addressing of these factors are crucial for the formulation of effective marketing strategies that resonate with industry collaborators [25], [44]. This component completes the answer to RQ4: intelligent marketing mitigates cultural and informational asymmetries by leveraging AI-driven profiling and adaptive communication tools that match academic outputs to the absorptive capacity and readiness level of each specific industrial partner.

The following Table VII synthesizes the university-industry collaboration process in the systematic literature review.

D. Integration and Interaction Mechanisms

This subsection further addresses RQ1 and RQ2 by specifying how each framework component interacts with the others to produce the operational conditions for effective AI-enhanced technology transfer. Kowledge Transfer Mechanisms represent the fundamental processes through which technological advancements and scholarly expertise from universities are disseminated to industry collaborators [7], [20], [15]. Patent licensing establishes formal frameworks for intellectual property transfer, spin-off enterprises facilitate extensive technology commercialization, consulting relationships promote

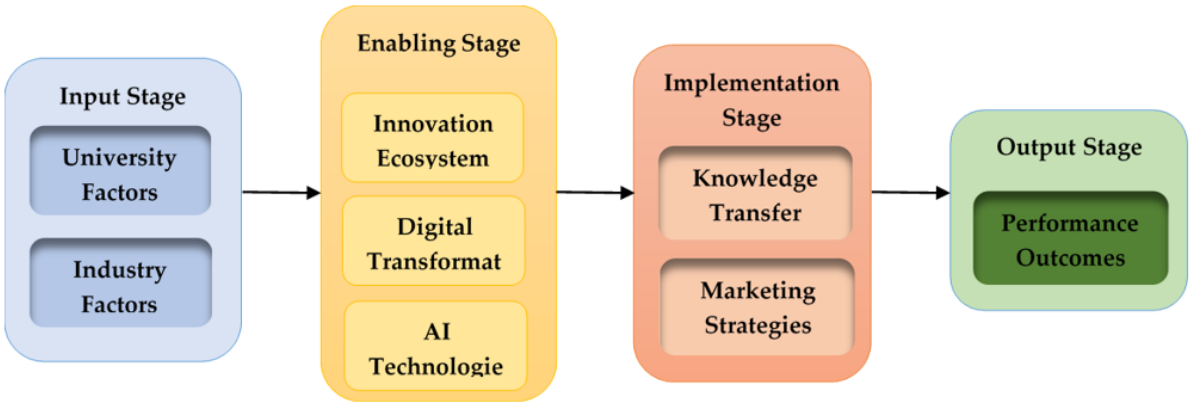


Fig. 4. Conceptual framework of AI-enhanced marketing for university-industry technology transfer.

TABLE VII. UNIVERSITY-INDUSTRY COLLABORATION PROCESS

Stage	Primary Activities	AI Enhancement Opportunities
Research Development	Basic and applied research, Innovation identification	Research trend analysis, Impact prediction
Technology Assessment	Feasibility analysis, Commercial potential evaluation	Automated assessment tools, Market potential algorithms
Market Analysis	Industry needs identification, Competitive landscape analysis	Real-time market intelligence, Predictive trend analysis
Partner Identification	Industry partner screening, Capability matching	Intelligent partner matching, Compatibility algorithms
Collaboration Framework	Agreement negotiation, Partnership structuring	Contract optimization, Risk assessment tools
Knowledge Transfer	Technology transmission, Expertise sharing	Personalized knowledge delivery, Learning optimization
Commercialization	Product development, Market launch	Launch strategy optimization, Success prediction
Performance Evaluation	Impact assessment, Success measurement	Automated performance tracking, Continuous improvement

direct expertise dissemination, collaborative research endeavors foster mutual innovation progress, and proof-of-concept facilities connect laboratory-based research with commercial viability [17], [18], [24]. These mechanisms must be augmented with artificial intelligence capabilities to enhance the efficiency, effectiveness, and scalability of technology transfer operations [19], [32]. Digital Transformation Elements offer the requisite technological framework and procedural modernization essential for the execution of AI-enhanced marketing initiatives [12], [31]. The digital infrastructure underpins data acquisition, analytical processes, and communication efforts, while data analytics competencies enable informed decision-making and performance enhancement [16], [25], [47]. Virtual platforms encourage online collaboration and relationship management, collaborative tools support distributed teamwork and knowledge exchange, and cloud computing technologies offer scalable and accessible computing resources [5], [42], [35]. Collectively, these elements establish a robust foundation for the implementation of advanced AI marketing strategies within technology transfer paradigms [41], [39]. This directly responds to RQ5: the maturity level of an institution's digital infrastructure — spanning cloud computing adoption, data analytics capacity, and virtual collaboration readiness — constitutes a critical contextual determinant of AI adoption success in research valorization. The Innovation Ecosystem encompasses the wider contextual factors that significantly affect the success of university-industry collaborations [22], [38]. Networks foster connections and opportunities for relationship development, partnerships establish formal collaboration frameworks, incubators assist in the advancement of nascent technologies, governmental support impacts policy and funding landscapes, and financial mechanisms dictate the availability of resources for collaborative ventures [6], [40], [44]. Adopting an ecosystem perspective acknowledges that effective technology transfer necessitates conducive environmental conditions that go beyond the capabilities of individual organizations [46], [45]. This extends the answer to RQ5 by situating AI adoption within an ecosystem logic: the presence or absence of supportive policy, regional networks, and funding infrastructures determines whether AI-enhanced marketing initiatives can be scaled and sustained.

Performance Outcomes signify the quantifiable results and effects of AI-enhanced marketing strategies within technology transfer contexts [21], [41], [30]. Metrics for technology commercialization evaluate the success of translating university research into marketable applications, knowledge transfer efficiency assessments gauge the effectiveness of transfer mechanisms, innovation performance indicators monitor the broader impacts of collaborative efforts, and evaluations of competitive advantage ascertain the strategic value generated through university-industry alliances [18], [38]. These outcomes furnish crucial feedback for the ongoing refinement and optimization of AI-enhanced marketing strategies [19], [40], [39]. This constitutes the direct operational answer to RQ7: the effectiveness of AI-enhanced marketing on university-industry collaborations can be assessed through a multi-dimensional

indicator system encompassing commercialization conversion rates, partnership formation speed, knowledge transfer efficiency scores, and long-term innovation impact metrics. The proposed conceptual model addresses several critical gaps identified in the current literature while providing a comprehensive framework for understanding and implementing AI-enhanced marketing in university-industry technology transfer contexts. The following Table VIII elucidates the framework of marketing augmented by artificial intelligence within the context of technology transfer between universities and industries.

The amalgamation of artificial intelligence technologies with marketing methodologies signifies a notable divergence from conventional technology transfer paradigms, presenting prospects for augmented efficiency, refined targeting, and enhanced relationship management [3], [2], [1]. Nevertheless, effective execution necessitates meticulous consideration of the organizational, technological, and environmental elements that impact adoption and efficacy [36], [13], [47].

E. Theoretical Contributions

This subsection addresses RQ6 from a theoretical standpoint by articulating the specific conceptual innovations introduced by the proposed framework in relation to the existing literature. The theoretical framework enhances the discourse on technology transfer by presenting the inaugural comprehensive amalgamation of artificial intelligence technologies, marketing methodologies, and mechanisms for knowledge transfer within a singular conceptual paradigm [20], [33], [29]. Prior investigations have predominantly examined these areas in isolation, which constrains the comprehension of their interrelated dynamics and synergistic capabilities [8], [7], [12]. The proposed framework elucidates how AI competencies can augment conventional marketing tactics while concurrently enhancing the efficacy of knowledge transfer via intelligent automation and data-informed decision-making [27], [14], [25]. These contributions collectively represent a direct and substantive answer to RQ6, establishing the conceptual architecture within which AI, marketing, and technology transfer can be coherently integrated.

Furthermore, the model propels the theory of digital transformation by delineating the technological and organizational prerequisites for the successful application of AI within academic settings [5], [31]. In contrast to generic frameworks of digital transformation, this model explicitly addresses the distinctive attributes of university-industry collaboration, which encompasses the significance of tacit knowledge transfer, the function of intellectual property management, and the impact of innovation ecosystem elements [7], [17], [19]. This specificity amplifies the practical significance and applicability of the framework for practitioners involved in technology transfer [15], [16], [32].

F. Practical Implications

This subsection primarily addresses RQ4 and RQ5 from a practitioner perspective. It specifies how intelligent marketing — operationalized through the proposed framework — can practically bridge the cultural and informational divide between academic and industrial actors, and what contextual investments are needed to enable AI adoption. For university administrators and professionals involved in technology transfer, the proposed framework offers practical directives for cultivating AI-driven marketing proficiencies that enhance industrial engagement and collaborative outcomes [4], [6], [18]. This model delineates particular technological investments, organizational capabilities, and strategic methodologies that institutions of higher education ought to prioritize to augment their effectiveness in technology transfer [20], [24]. Strategies for implementation should concentrate on fostering AI competencies within technology transfer offices, establishing digital infrastructures for data analytics and relationship management, and devising comprehensive training initiatives aimed at staff development [5], [12], [47]. Industry collaborators can employ the framework to gain insights into university capabilities and opportunities for collaboration while simultaneously enhancing their absorptive capacity for novel technologies [17], [30]. The model underscores the significance of forging dedicated partnerships with universities to facilitate AI-enhanced collaboration, investing in digital platforms to engage with universities, and cultivating internal capabilities necessary for assessing and implementing university technologies [18], [19], [42]. Such investments empower industry partners to optimize the benefits derived from university-industry collaboration while also contributing to the development of innovation ecosystems [15], [38]. Policymakers can utilize the framework to formulate conducive regulations, funding strategies, and infrastructural investments that promote AI-enhanced technology transfer [22], [40], [44]. The model identifies essential success factors warranting policy focus, including the advancement of AI infrastructure, regulatory frameworks that facilitate digital collaboration, and incentive structures that foster university-industry partnership development [16], [46], [41]. Policy interventions must tackle both technological and organizational obstacles while advocating for equitable access to capabilities associated with AI-enhanced technology transfer [25], [39], [45].

G. Limitations and Considerations

This subsection is relevant to RQ7, as the limitations acknowledged here define the boundary conditions under which the proposed performance indicators can be reliably applied. It also partially addresses RQ5 by acknowledging that contextual variability across institutional settings limits the generalizability of AI adoption prescriptions. The proposed framework is predicated on a synthesis of existing literature rather than empirical validation, thereby necessitating subsequent investigations to evaluate the model's efficacy in practical contexts [21], [37]. The swift advancement of artificial intelligence technologies may necessitate regular revisions to the framework to ensure its ongoing relevance and precision [2], [23]. Furthermore, the model may require modifications to suit diverse institutional frameworks, cultural landscapes, and industrial domains that present distinct characteristics and demands [13], [38], [40]. Challenges in implementation encompass the considerable resource allocations essential for the development of AI infrastructure, the necessity for specialized technical competencies that may be scarce within academic institutions, and potential resistance from stakeholders who are accustomed to conventional technology transfer methodologies [10], [36]. Organizations must meticulously evaluate their preparedness

TABLE VIII. CONCEPT OF AI-ENHANCED MARKETING FRAMEWORK FOR UNIVERSITY-INDUSTRY TECHNOLOGY TRANSFER

Theme	Key Concepts	Research Gaps	Future Directions
University-Industry Technology Transfer	Patent licensing; Spin-offs; Collaboration barriers; Technology commercialization	Limited AI integration in TT processes; Need for digital marketing strategies	AI-enhanced TT processes; Digital collaboration platforms; Smart matching systems
AI in Marketing Strategies	Personalization; Predictive analytics; Automation; Data-driven strategies	Insufficient application in TT context; Limited university-industry focus	AI-powered university marketing; Intelligent partnership platforms; Predictive TT analytics
Digital Transformation	Digital infrastructure; Virtual collaboration; Cloud technologies; Online platforms	Lack of comprehensive frameworks; Missing integration with marketing strategies	Comprehensive digital TT ecosystems; AI-driven transformation models
Knowledge Management	Knowledge sharing; Tacit knowledge transfer; Organizational learning; Absorptive capacity	Underexplored AI enhancement; Limited digital transformation integration	AI-enhanced knowledge discovery; Intelligent knowledge matching systems
Innovation Ecosystems	Networks; Partnerships; Incubators; Government support; Funding mechanisms	Missing AI-enhanced networking; Insufficient digital ecosystem models	Smart innovation networks; AI-powered ecosystem orchestration
Conceptual Models & Frameworks	Theoretical frameworks; Structural models; Process models; Integration models	Absence of AI-marketing-TT integration models; Limited comprehensive frameworks	Integrated AI-marketing-TT models; Multi-dimensional frameworks

for AI integration while formulating comprehensive change management frameworks that cater to both technological and cultural transformation imperatives [31], [47], [35].

The literature points out both the benefits and the limitations of using AI marketing in traditional university-Industry technology transfer (UITT) models. Many argue that the digital transformation and the use of artificial intelligence (AI) can contribute to better collaboration, but only if the institution has an adequate digital infrastructure and is able to absorb new knowledge. In the absence of such characteristics, collaboration may not improve, especially with an internal resistance to change. Some cite the over-dependence on automated tasks, which may eliminate trust and the dynamics necessary for university–industry collaboration.

The existing literature reflects major differences between developed and emerging markets in terms of the level of technology, support for artificial intelligence (AI) integration, and funding. While some authors describe improved matching of partners, the efficiency of communication, and results of the commercialization process, increased AI integration can require the complexity of the organization and substantial resource investments with no guaranteed return on the investment. As a result, there are significant operational contradictions in AI marketing in UITT models. These inconsistencies reinforce the requirement for an integrative and context-sensitive conceptual framework.

H. Future Research Directions

This subsection most directly addresses RQ7 by specifying the empirical and methodological approaches needed to develop and validate the performance indicators through which the effectiveness of AI-enhanced marketing on university-industry collaborations can be rigorously assessed. It also opens new avenues for RQ2 and RQ5 by identifying algorithmic and contextual research priorities. The proposed theoretical framework lays the groundwork for extensive future investigations across various disciplines and methodological paradigms. Empirical validation initiatives represent the most pressing research focus, necessitating systematic evaluations of the framework's efficacy within diverse university-industry collaboration environments [21], [41]. Such studies ought to utilize stringent experimental or quasi-experimental methodologies that can elucidate causal links between AI-enhanced marketing interventions and technology transfer results while effectively controlling for extraneous variables [19], [38], [37]. Longitudinal research methodologies possess particular significance for comprehending the dynamic interactions among the components of the framework and their progression over time [37]. Such investigations can monitor the advancement of AI capabilities, the evolution of marketing strategies, and the enduring effects of technology transfer programs while pinpointing essential success determinants and potential pitfalls [18], [46]. Cross-sectional comparative analysis may scrutinize the framework's effectiveness across various university classifications, industry categories, and national settings to elucidate boundary conditions and necessary adaptations [13], [40], [39].

Methodological advancements should investigate the amalgamation of sophisticated analytical techniques, including network analysis for deciphering collaboration patterns, machine learning methodologies for forecasting partnership success, and natural language processing for evaluating communication efficacy [30], [42]. Mixed-methods research frameworks can integrate quantitative outcome assessments with qualitative insights regarding stakeholder experiences, organizational dynamics, and cultural elements that affect implementation success [20], [37].

Research on technology development should concentrate on the creation of specialized AI instruments and platforms tailored explicitly for university-industry technology transfer applications [27], [14], [25]. These innovations may encompass intelligent matching systems that align university technologies with industry demands, predictive analytics platforms that estimate collaboration success probabilities, and automated relationship management systems that enhance stakeholder engagement [19], [23]. Such technological advancements necessitate close cooperation among computer science scholars,

marketing professionals, and technology transfer specialists [12], [17].

In direct response to RQ7, future research should prioritize the operationalization and validation of the Performance Outcomes component (Component 8), developing standardized indicator sets — including composite AI-effectiveness scores, collaboration conversion dashboards, and longitudinal innovation impact indices — that can be applied across diverse institutional and geographic contexts to provide replicable, benchmarkable measures of AI-enhanced marketing effectiveness.

VI. Conclusion

This integrative literature review and conceptual model development investigation addresses a significant deficiency in understanding how artificial intelligence can enhance marketing strategies for transferring technology knowledge between universities and industries. The proposed AI-enhanced marketing framework provides a comprehensive theoretical foundation that encompasses eight interrelated components essential for the effective implementation of technology transfer initiatives. The framework illustrates the potential of AI technologies to revolutionize conventional marketing methodologies while concurrently enhancing the efficacy of knowledge transfer through intelligent automation, data-informed decision-making, and improved stakeholder involvement.

The literature review uncovers substantial prospects for progressing both theoretical insights and practical applications of AI-enhanced marketing within university-industry collaborative environments. Nonetheless, the successful execution necessitates meticulous consideration of organizational preparedness, technological infrastructure, and ecosystem-level support factors that affect adoption and overall effectiveness. The identification of research voids and prospective avenues delineates a pathway for scholars and practitioners endeavoring to advance this nascent domain, while tackling pivotal challenges in the efficacy of technology transfer.

The practical ramifications of this inquiry extend beyond singular organizations to include policy recommendations, strategies for ecosystem development, and frameworks for international collaboration that can bolster global innovation capacity. As AI technologies continue to progress and mature, the amalgamation of these capabilities with marketing strategies signifies a transformative shift in how universities interact with industry partners and commercialize research outputs. Future research and implementation initiatives should prioritize empirical validation, technological advancement, and practical directives that facilitate the widespread adoption of AI-enhanced marketing strategies in contexts of technology transfer.

References

[1] C. Nuță, "Artificial intelligence and marketing strategies. A bibliometric perspective," *MANAGEMENT STUDENT WORKING PAPERS*, vol. 2, 2024. https://doi.org/10.24818/MSWP.2024.07

[2] M. Potwora, O. Vdovichena, D. Semchuk, L. Lipych, and V. Saienko, "The Use of Artificial Intelligence in Marketing Strategies: Automation, Personalization and Forecasting," *Journal of Management World*, vol. 2024, no. 2, Art. 2, 2024. https://doi.org/10.53935/jomw.v2024i2.275

[3] M. T. Nuseir and G. E. Refae, "The role of artificial intelligence, marketing strategies, and organizational capabilities in organizational performance: The moderating role of organizational behavior," *Uncertain Supply Chain Management*, vol. 10, no. 4, pp. 1457–1466, 2022. https://doi.org/10.5267/j.uscm.2022.6.010

[4] M. Blankesteijn, B. Bossink, and P. Van Der Sijde, "Science-based entrepreneurship education as a means for university-industry technology transfer," *International Entrepreneurship and Management Journal*, vol. 17, no. 2, pp. 779–808, 2021. https://doi.org/10.1007/s11365-019-00623-3

[5] D. Rico-Bautista, C. D. Guerrero, C. A. Collazos, G. Maestre-Góngora, J. A. Hurtado-Alegría, Y. Medina-Cárdenas, and J. Swaminathan, "Smart University: A vision of technology adoption," *Revista Colombiana de Computación*, vol. 22, no. 1, Art. 1, 2021. https://doi.org/10.29375/25392115.4153

[6] R. Ravi and M. D. Janodia, "University-Industry Technology Transfer in India: A Plausible Model Based on Success Stories from the USA, Japan, and Israel," *Journal of the Knowledge Economy*, vol. 13, no. 2, pp. 1692–1713, 2022. https://doi.org/10.1007/s13132-022-00908-z

[7] P. Lee, "Chapter 10 Tacit knowledge and university-industry technology transfer," 2020. https://china.elgaronline.com/edcollchap/edcoll/9781788116626/9781788116626.00020.xml

[8] D. Battaglia, E. Paolucci, and E. Ughetto, "Hurdles in University-Industry Technology Transfer: Why Research-Based Inventions are Not Transferred to the Market?", *IEEE Transactions on Engineering Management*, vol. 71, pp. 13415–13427, 2024. https://doi.org/10.1109/TEM.2023.3269731

[9] A. D. Daniel and L. Alves, "University-industry technology transfer: The commercialization of university's patents," *Knowledge Management Research & Practice*, vol. 18, no. 3, pp. 276–296, 2020. https://doi.org/10.1080/14778238.2019.1638741

[10] S. Ayesha and D. S. Pougajendy, "A Study on the Applicability of Artificial Intelligence Marketing for Creating Data-Driven Marketing Strategies at Cherri Technologies, Pondicherry," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 6, no. 6, 2024.

[11] F. Qasim Hasan, "Impact of Artificial Intelligence in Companies Marketing Strategies (UNDER THE GUIDANCE No. 22GSOB2011035)," School of Business, Galgotias University, pp. 1–5, 2024. https://ijsrem.com/download/impact-of-artificial-intelligence-in-companies-marketing-strategies/

[12] N. S. Makrydakis, "Digital Darwinism: Digital transformation, marketing and public higher education in Greece," *IROCAMM – International Review Of Communication And Marketing Mix*, vol. 5, no. 1, pp. 21–35, 2022. https://dx.doi.org/10.12795/IROCAMM.2021.v05.i01.02

[13] M. U. Mubarok, M. I. Sari, Y. G. Wibowo, and R. Mathew, "Comparative Study of Artificial Intelligence (AI) Utilization in Digital Marketing Strategies Between Developed and Developing Countries: A Systematic Literature Review," *Ilomata International Journal of Management*, vol. 6, no. 1, Art. 1, 2025. https://doi.org/10.61194/ijjm.v6i1.1534

[14] J. Tauheed, A. Shabbir, and M. S. Pervez, "Exploring the Role of Artificial Intelligence in Digital Marketing Strategies," *Journal of Business, Communication & Technology*, pp. 54–65, 2024. https://doi.org/10.56632/bct.2024.3105

[15] J. Eidlisz, I. von Simson, and G. Gold-von Simson, "Exploring the current state of technology transfer in the United States: Perspectives and improvement strategies from the experts," *Frontiers in Research Metrics and Analytics*, vol. 9, Art. 1376185, 2024. https://doi.org/10.3389/frma.2024.1376185

[16] A. Ahmad, R. Ramli, and F. Sh. Ahmad, "Fostering Global Branding Prominence of Public Higher Education Institutions in Brunei Darussalam through Strategic Resources of Global Collaboration and Digital Transformation," *International Journal of Academic Research in Business and Social Sciences*, vol. 14, no. 10, pp. 3157–3162, 2024. https://doi.org/10.6007/IJARBSS/v14-i10/23470

[17] A. Yazdani, "Developing a Knowledge Management Framework for Innovation Performance in Tech Startups," *Business, Marketing, and Finance Open*, vol. 1, no. 3, Art. 3, 2024. https://doi.org/10.61838/bmfopen.1.3.5

[18] H. Hu, D. Zhang, and C. Wang, "Impact of social media influencers' endorsement on application adoption: A trust transfer perspective," *Social Behavior and Personality: An International Journal*, vol. 47, no. 11, pp. 1–12, 2019. https://doi.org/10.2224/sbp.8518

[19] S. Hadianfar, M. Malekinia, M. Khalaj, and A. Shahmansoury, "Presenting and Validating a Digital Transformation Model in Iran's Shahrvand Chain Stores," *Business, Marketing, and Finance Open*, vol. 1, no. 3, Art. 3, 2024. https://doi.org/10.61838/bmfopen.1.3.1

[20] V. V. Sviatetskyi, "Building A Management Framework For Intellectual Property Commercialisation," *Bulletin of the Kyiv National University of Technologies and Design. Series: Economic sciences*, vol. 153, no. 6, pp. 116–121, 2020. https://doi.org/10.30857/2413-0117.2020.6.13

[21] P. Singh, L. Arora, and A. Choudhry, "Consumer Behavior in the Service Industry: An Integrative Literature Review and Research Agenda," *Sustainability*, vol. 15, no. 1, Art. 1, 2023. https://doi.org/10.3390/su15010250

[22] O. Fadeyi, P. Maresova, R. Stemberkova, M. Afolayan, and F. Adeoye, "Perspectives of University-Industry Technology Transfer in African Emerging Economies: Evaluating the Nigerian Scenario via a Data Envelopment Approach," *Social Sciences*, vol. 8, no. 10, Art. 10, 2019. https://doi.org/10.3390/socsci8100286

[23] O. Kotyrlo, R. Naboka, V. Nestor, D. Tyshko, and O. Panasenko, "Ways to use artificial intelligence to improve the personalization of marketing strategies and improve the effectiveness of communication with consumers," *Multidisciplinary Reviews*, vol. 8, Art. 2024spe074, 2024. https://doi.org/10.31893/multirev.2024spe074

[24] F. Shao and J. Ai, "An Exploratory Theoretical Framework based on the Interactive Marketing Model of Industrial Products based on user Innovation," pp. 428–432, 2017. https://doi.org/10.2991/meici-17.2017.81

[25] M. I. Abubakirova, "Development of Library Network Resources in the Context of Digital Transformation (marketing Approach)," pp. 6–16, 2022. https://doi.org/10.25206/978-5-8149-3568-7-2022-6-16

[26] R. Daniel and C. Alves, "University-Industry Collaboration: Mechanisms and Models for Knowledge Transfer," *Journal of Technology Transfer*, vol. 45, pp. 1452–1471, 2020. https://doi.org/10.1007/s10961-020-09769-4

[27] I. Gabelaia, "The Applicability of Artificial Intelligence Marketing for Creating Data-driven Marketing Strategies," *Journal of Marketing Research and Case Studies*, pp. 1–11, 2024. https://doi.org/10.5171/2022.466404

[28] J. F. S. A. Ghezzi, E. de F. R. Higa, M. A. Lemes, and M. J. S. Marin, "Strategies of active learning methodologies in nursing education: An integrative literature review," *Revista Brasileira de Enfermagem*, vol. 74, Art. e20200130, 2021. https://doi.org/10.1590/0034-7167-2020-0130

[29] C. Gong and V. Ribiere, "Understanding the role of organizational agility in the context of digital transformation: An integrative literature review," *VINE Journal of Information and Knowledge Management Systems*, vol. 55, no. 2, pp. 351–378, 2025. https://doi.org/10.1108/VJIKMS-09-2022-0312

[30] J. Jiang, Y. Zhao, and J. Feng, "University-Industry Technology Transfer: Empirical Findings from Chinese Industrial Firms," *Sustainability*, vol. 14, no. 15, Art. 15, 2022. https://doi.org/10.3390/su14159582

[31] C. Lukita, M. Wali, R. A. Sunarjo, E. P. Harahap, F. P. Oganda, and N. Khofifah, "Digital Transformation in Reading Behavior: A Business Model Canvas Approach for Marketing E-Books Library," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–8, 2023. https://doi.org/10.1109/CITSM60085.2023.10455122

[32] B. Marks and J. Thomas, "Adoption of virtual reality technology in higher education: An evaluation of five teaching semesters in a purpose-designed laboratory," *Education and Information Technologies*, vol. 27, no. 1, pp. 1287–1305, 2022. https://doi.org/10.1007/s10639-021-10653-6

[33] V. C. Meireles, C. M. G. Labegalini, and V. D. A. Baldissera, "Tracer Methodology and the quality of care: Integrative literature review," *Rev Gaúcha Enferm*, 2019. https://doi.org/10.1590/1983-1447.2019.20180142

[34] L. A. Nguyen, R. Evan, S. Chaudhuri, M. Hagen, and D. Williams, "Inclusion in the workplace: An integrative literature review," *European Journal of Training and Development*, vol. 48, no. 3/4, pp. 334–356, 2023. https://doi.org/10.1108/EJTD-10-2022-0104

[35] D. Opoku, F. Pobee, and R. O. Okyireh, "Determinants of E-Learning System Adoption among Ghanaian University Lecturers: An Application of Information System Success and Technology Acceptance Models," *American Journal of Social Sciences and Humanities*, vol. 5, no. 1, Art. 1, 2020. https://doi.org/10.20448/801.51.151.168

[36] F. Qasim Hasan, "Impact of Artificial Intelligence in Companies Marketing Strategies (UNDER THE GUIDANCE No. 22GSOB2011035)," School of Business, Galgotias University, pp. 1–5, 2024. https://ijsrem.com/download/impact-of-artificial-intelligence-in-companies-marketing-strategies/

[37] S. Rahimi, A. Khorasani, M. Rezaeizadeh, and J. Waterworth, "The virtual human resource development (VHRD) approach: An integrative literature review," *European Journal of Training and Development*, vol. 46, no. 5/6, pp. 484–503, 2021. https://doi.org/10.1108/EJTD-08-2020-0131

[38] A. Rejeb, K. Rejeb, A. Appolloni, Y. Kayikci, and M. Iranmanesh, "The landscape of public procurement research: A bibliometric analysis and topic modeling based on Scopus," *Journal of Public Procurement*, vol. 23, no. 2, pp. 145–178, 2023. https://doi.org/10.1108/JOPP-06-2022-0031

[39] C. A. H. Salazar, V. M. V. Sir, F. T. C. Zubiate, E. G. Paco, S. E. R. Huamán, and S. C. C. Zubiate, "UNTRM On the Global Scene: A Bibliometric Analysis of Its Scientific Production Indexed in Scopus," *Journal of Ecohumanism*, vol. 3, no. 3, Art. 3, 2024. https://doi.org/10.62754/joe.v3i3.3397

[40] V. Sánchez-Castillo, "Analysis of Colombian scientific production around Agroecology in Scopus," *Data and Metadata*, vol. 3, pp. 618–618, 2024. https://doi.org/10.56294/dm2024.618

[41] K. Shahzad, S. A. Khan, and A. Iqbal, "Effects of artificial intelligence on university libraries: An SLR of cite score and IF journals' articles from 2018 to 2023," *Global Knowledge, Memory, and Communication*, ahead-of-print(ahead-of-print), 2024. https://doi.org/10.1108/GKMC-12-2023-0498

[42] C. E. Smith, K. Shiekh, H. Cooreman, S. Rahman, Y. Zhu, M. K. Siam, M. Ivanitskiy, A. M. Ahmed, M. Hallinan, A. Grisak, and G. Fierro, "Early Adoption of Generative Artificial Intelligence in Computing Education: Emergent Student Use Cases and Perspectives in 2023," *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V.1*, pp. 3–9, 2024. https://doi.org/10.1145/3649217.3653575

[43] J. Thomas and A. Harden, "Methods for the thematic synthesis of qualitative research in systematic reviews," *BMC Med Res Methodol*, vol. 8, p. 45, 2008. https://doi.org/10.1186/1471-2288-8-45

[44] O. Velychko, L. Velychko, and Y. Yahnin, "Logistical Models Of Competitive Development In Conditions Of The Digital Transformation In Production And Marketing Business Processes," *Economy and State*, vol. 3, pp. 40–45, 2021.

[45] N. N. S. P. Verawati and A. Ramdani, "Research Trend of Local Wisdom Issues Based on Scopus Journal Database: A Bibliometric Study," *International Journal of Contextual Science Education*, vol. 1, no. 1, pp. 11–21, 2023. https://doi.org/10.29303/ijcse.v1i1.551

[46] M. Yonfá-Medranda, D. Sabando-Vera, J. Pérez-Gálvez, and C. Jara-Alba, "Consumer ethnocentrism: Bibliometric analysis and literature review through cognitive structure and mapping of research based on a web of science (WoS) and Scopus," *Cogent Business & Management*, vol. 11, no. 1, Art. 2377320, 2024. https://doi.org/10.1080/23311975.2024.2377320

[47] A. Zarifis and L. Efthymiou, "The four business models for AI adoption in education: Giving leaders a destination for the digital transformation journey," in *2022 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1868–1872, 2022. https://doi.org/10.1109/EDUCON52537.2022.9766687

[48] Z. Zuckerman, Z. Nuryana, A. Pangarso, and F. M. Zain, "Factor of Zoom cloud meetings: Technology adoption in the pandemic of COVID-19," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 10, no. 3, Art. 3, 2021. https://doi.org/10.11591/ijere.v10i3.21726

Development of an Arabic Pet Adoption System with Hybrid Recommendations

Hailah Alballaa
Information Technology Department-College of Computer and Information Sciences,
King Saud University, Riyadh, Saudi Arabia

***Abstract*—Pet adoption platforms often face challenges in effectively matching adopters with suitable pets due to limited personalization and the lack of localized language support. This study presents EWAA, an Arabic-enabled mobile platform for pet adoption in Saudi Arabia that integrates a hybrid recommender system combining content-based and collaborative filtering techniques. The content-based component constructs a user profile from one-hot encoded pet attributes (category, breed, color, and age) of previously liked pets, while the collaborative filtering component identifies similar users through cosine similarity and recommends pets based on their preferences. The two components operate independently in a mixed hybrid configuration, which mitigates the cold-start limitation of collaborative filtering and the over-specialization limitation of content-based filtering. The platform also emphasizes usability, accessibility, and privacy for both adopters and pet owners through a fully Arabic interface and controlled information visibility. The system was evaluated with 20 questionnaire participants and 6 interview participants using User Acceptance Testing (UAT) and Non-Functional Requirements (NFR) testing. Results indicate high levels of user satisfaction, with task completion rates of 100% on 10 of 11 test scenarios, page load times between 1 and 12 seconds, and learning times between 1 and 8 minutes, suggesting that the proposed approach provides a viable foundation for supporting the pet adoption process in Arabic-speaking contexts.**

Keywords*—Recommender systems; pet adoption; mobile applications; collaborative filtering; content-based filtering; user experience

I. Introduction

Pet adoption plays an important role in improving animal welfare while providing companionship and quality-of-life benefits to individuals and families [1]. Despite these benefits, the adoption process remains fragmented in many contexts, relying on physical shelter visits, disconnected platforms, or informal social media channels, which can be inefficient and time-consuming for prospective adopters.

With the growing interest in pet adoption in Saudi Arabia, there is an increasing demand for digital platforms that provide localized and accessible services. However, existing pet adoption platforms exhibit two notable limitations in this context. First, they lack Arabic language support, creating usability barriers for native Arabic speakers. Second, they typically rely on either content-based or collaborative filtering in isolation, each of which suffers from well-known weaknesses: content-based filtering is prone to over-specialization, while collaborative filtering suffers from the cold-start problem [2].

This study presents EWAA, an Arabic-enabled mobile pet adoption platform that integrates a hybrid recommender approach combining content-based and collaborative filtering techniques. The objectives of this work are: first, to design and implement an Arabic-language mobile adoption platform tailored to the Saudi context; second, to integrate a mixed hybrid recommender that mitigates the cold-start and over-specialization limitations of single-method approaches; and third, to evaluate the resulting system through user acceptance and non-functional requirements testing. The significance of the work lies in providing a deployed adoption platform that combines hybrid recommendation with native Arabic language support, addressing both a regional accessibility gap and a methodological gap in prior pet adoption recommenders. The study focuses on cats and dogs, which represent the majority of adoption cases in the target context.

The remainder of this study is organized as follows: Section II reviews related work on pet adoption platforms and recommender approaches. Section III presents the system overview, and Section IV describes the system architecture. Section V details the implementation, and Section VI presents the design of the hybrid recommender. Section VII reports the system evaluation, and Section VIII concludes the study and outlines directions for future work.

II. Related Work

Recommender systems are well established as a means of mitigating information overload and supporting personalized decision-making [3]. They are commonly categorized into content-based, collaborative, and hybrid approaches, with hybrid configurations typically introduced to mitigate the cold-start and over-specialization limitations of the individual methods [2]. Within the pet adoption domain, several lines of work are particularly relevant.

Karani et al. [4] developed a structured dataset and a recommender system to suggest appropriate canine and feline breeds to prospective adopters. Their work provides a foundation for breed-specific recommendations in pet adoption systems. However, the system operates at the level of breeds rather than individual pets, which limits its ability to differentiate between specific animals of the same breed in a real adoption listing, and it does not incorporate user interaction signals such as likes or other implicit feedback.

De Luna et al. [5] compared content-based filtering, collaborative filtering, and k-nearest neighbors for dog breed recommendation, and reported content-based filtering as the best-performing approach.

Montiveles et al. [6] used item-based collaborative filtering with Pearson correlation and k-nearest neighbors to recommend adoptable pets. As a collaborative only approach, the

system is susceptible to the cold-start limitation when new users or pets have limited interaction history, and it does not incorporate content-based signals from pet attributes that could mitigate this gap.

Chaudhari et al. [7] presented a recommender system that uses content-based filtering with cosine similarity metrics to match potential adopters with suitable pets, taking into account various factors such as breed characteristics, adopter lifestyle, and living conditions, and offering a holistic approach to the adoption matching process. However, as a purely content-based method it inherits the well-known over-specialization problem [2]. The approach also does not address the cold-start case in which the active user has not yet liked any pet.

Across these studies, two gaps are relevant to the present work. First, none of the prior systems is evaluated or deployed in an Arabic-language context. Second, with the exception of the comparative evaluation in [5], none of them combines content-based and collaborative signals within a single deployed adoption system, despite the fact that hybrid configurations are a standard remedy for the cold-start and over-specialization limitations identified above. EWAA addresses these gaps by adopting a mixed hybrid configuration [2] over user-supplied pet attributes and interaction data, within a fully Arabic-language mobile platform.

Beyond research prototypes, several deployed platforms support pet adoption services and provide a useful basis for identifying the functional capabilities expected in a mature adoption application. These platforms include PetMatch [8], Petfinder [9], and Pet Adoption [10]. As shown in Table I, a comparison of these platforms highlights their key features and limitations.

The EWAA application addresses several limitations present in existing platforms by offering a user-centric experience tailored to the needs of adopters in Saudi Arabia. One of its key features is Arabic language support, which ensures a seamless navigation experience and improves accessibility for local users. Additionally, EWAA enhances privacy protection by keeping the personal details of pet owners and adopters confidential until an adoption request is accepted, fostering a secure and trustworthy adoption environment. The application also features an intuitive and engaging interface, designed to be visually appealing and user-friendly, which encourages greater interaction and simplifies the pet adoption process. By integrating these improvements, EWAA improves usability, security, and user engagement, providing an optimized platform for pet adoption in the region.

The comparative analysis presented in Table I highlights the strengths of the EWAA system compared to existing pet adoption platforms. Although current systems provide basic features such as browsing, filtering, and pet listing, they lack key features such as Arabic language support, integrated privacy protection, and personalization. In contrast, EWAA combines these capabilities within a single platform by offering accessibility, personalized pet recommendations, and secure adoption workflows. This integration enhances the user experience and supports a more efficient and user-centered pet adoption process.

TABLE I. COMPARISON OF KEY FEATURES BETWEEN EXISTING PET ADOPTION PLATFORMS AND EWAA.

Feature	Pet Adoption	PetMatch	Petfinder	EWAA
User registration	No	No	Yes	Yes
Browse categories	No	No	Yes	Yes
Search and filtering	Yes	Yes	Yes	Yes
Submit request	No	No	Yes	Yes
Add pets	Yes	Yes	Yes	Yes
Recommendations	No	No	Yes	Yes
Favorites	No	Yes	Yes	Yes
Social sharing	Yes	No	Yes	Yes
Arabic support	No	No	No	Yes
Pet care section	No	No	Yes	Yes
User-friendly UI	No	No	Yes	Yes
Privacy protection	No	No	No	Yes

III. SYSTEM OVERVIEW

The proposed system, EWAA, is a mobile-based pet adoption platform designed to streamline and improve the adoption process in Saudi Arabia. It provides a centralized and user-friendly environment that connects potential adopters with pet owners, reducing reliance on traditional methods such as physical shelter visits and informal social media channels.

EWAA provides a unified user model in which a registered user can act as both a pet owner and an adopter depending on the system interaction. This design simplifies user management and enhances flexibility within the platform. Users can browse available pets, apply search and filtering options, and submit adoption requests. Pet owners, on the other hand, can add, update, and manage pet listings, as well as review and respond to adoption requests.

A key feature of the system is the integration of a hybrid recommender system, which combines content-based and collaborative filtering techniques to generate personalized pet suggestions. Recommendations are based on user preferences and interactions, such as selected pet attributes and previously liked pets, improving the likelihood of successful matches between adopters and pets.

In addition to core adoption functionalities, EWAA emphasizes usability and accessibility through its fully Arabic-supported interface, enabling a localized experience for Arabic-speaking users. The system also incorporates privacy mechanisms that ensure user information remains confidential until adoption requests are approved, thereby fostering a secure and trustworthy environment.

Overall, EWAA is designed to improve the efficiency, accessibility, and user experience of the pet adoption process through user-centered recommendation capabilities.

To further illustrate the interaction between users and the system, Fig. 1 presents the main use case diagram of the EWAA application. The diagram highlights the primary functionalities available to both adopters and pet owners, including browsing pets, submitting adoption requests, managing pet listings, and receiving recommendations.

IV. SYSTEM ARCHITECTURE

The system follows a client-server communication model, in which the client sends requests to the server, and the server

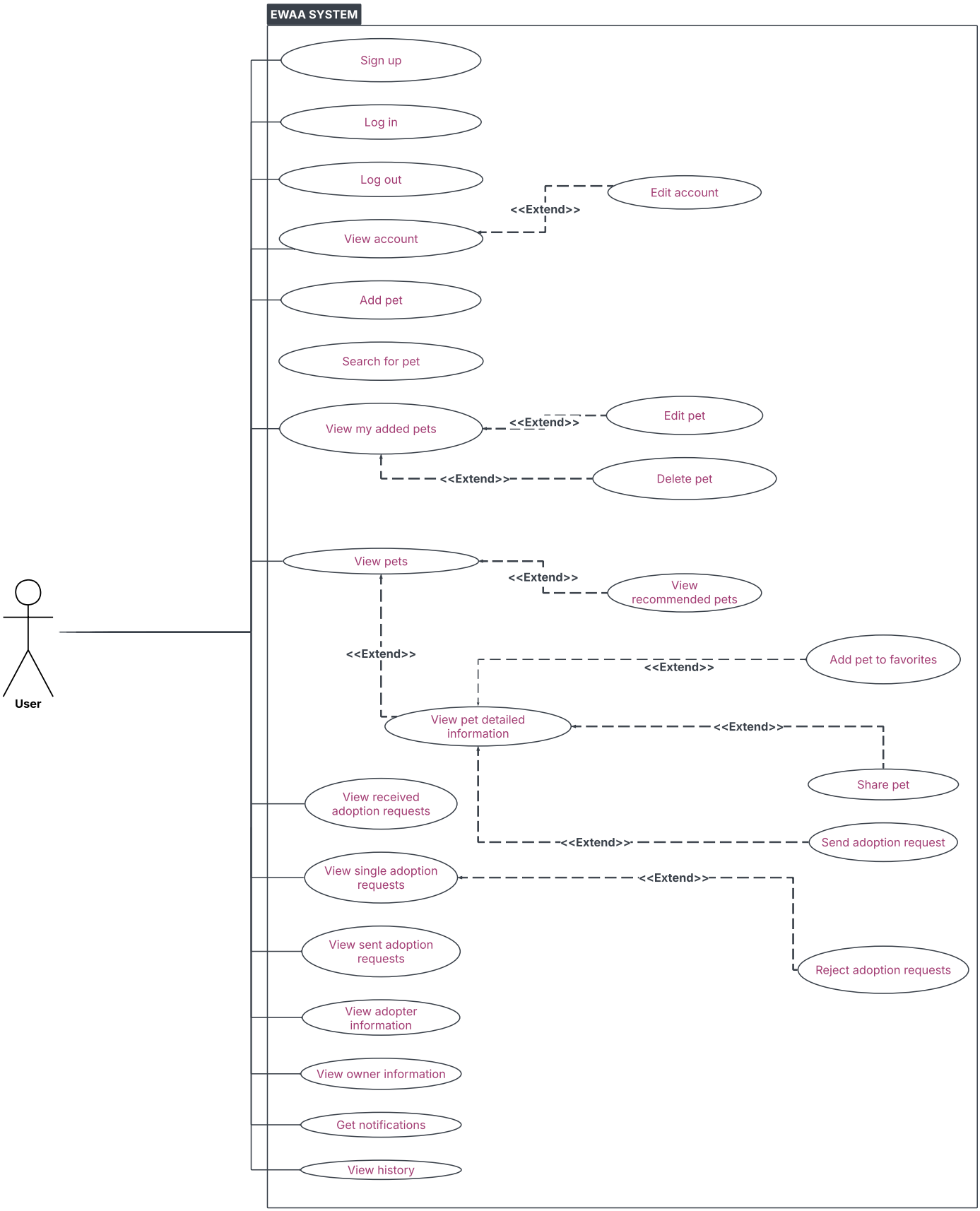


Fig. 1. Use case diagram of the EWAA system.

processes these requests and returns appropriate responses. This architecture supports centralized data management and enables users to access the application from different locations.

Fig. 2 illustrates the overall system architecture of EWAA. The client side is implemented as a mobile application developed using the Flutter framework and the Dart programming language. It provides users with an interactive interface to perform key actions such as browsing pets, submitting adoption requests, managing pet listings, and receiving recommendations. The server side is supported by a cloud-based NoSQL database, allowing efficient data retrieval and scalability.

A dedicated recommender system module, implemented in Python, is integrated into the architecture to generate personalized pet recommendations. The application collects user interaction data, such as preferences and liked pets, and sends this data to the backend database. The recommender module retrieves the stored data, applies the recommendation algorithms described in Section VI, and returns a list of suggested pets to the application, where they are displayed on the user's home page.

V. Implementation

A. Development Methodology

The EWAA system was developed using an agile methodology, specifically the Scrum framework [11], to ensure flexibility and iterative progress throughout the development process. Scrum enables the delivery of system functionalities in incremental stages through short development cycles known as sprints. Each sprint focused on implementing a set of user stories derived from the system requirements.

The development process included key Scrum practices such as sprint planning, progress tracking, and sprint review sessions. These iterative cycles allowed for continuous refinement of system features based on feedback and evolving user requirements. This approach ensured that the system remained aligned with user needs while supporting efficient development and timely delivery.

B. System Implementation

The EWAA system was implemented using a combination of modern development technologies to support scalability, performance, and usability. The mobile application was developed using the Flutter framework and the Dart programming language, providing a responsive and user-friendly interface.

Firebase was used as a Backend-as-a-Service (BaaS) platform, offering functionalities such as user authentication, real-time database management, and cloud storage. The system stores user data, pet information, and interaction records in a cloud-based NoSQL database.

The EWAA system includes a set of core features designed to enhance user experience and streamline the adoption process:

- User Registration and Authentication: Secure account creation, login, and email verification.
- Pet Management: Adding, editing, and managing pet listings with detailed attributes.
- Adoption Requests: Submitting, tracking, accepting, or rejecting adoption requests.
- Recommender System: Personalized suggestions using content-based and collaborative filtering.
- Notifications: Updates on adoption requests and activities related to pets.
- User Interface: Intuitive navigation with filtering and browsing capabilities.
- Privacy Protection: Secure handling of user data and controlled information visibility.

C. Data Collection

To support the recommender system, a data collection process was conducted focusing on cats and dogs, as they represent the most commonly adopted companion animals in Saudi Arabia. Initially, interviews with animal shelters were conducted to identify commonly adopted pet breeds in the region. Based on these findings, key characteristics of each breed were identified, including attributes such as energy level, affection, emotional reactivity, and intelligence.

Since the EWAA system was not deployed at the time of development, real user interaction data were not available. Therefore, a dataset was constructed by collecting information about the pets of real pet owners using an online questionnaire. The data collected included:

- Pet nickname
- Pet gender, age, and color
- Pet type and breed
- Health information (vaccination, sterilization, health profile)
- Personality traits

This dataset was used to initialize the recommender system and support personalized pet suggestions.

The system data structure is modeled, as shown in Fig. 3. The data model captures the main entities of the system — users, pets, adoption requests, and notifications — along with the relationships between them.

Each user can act as both an adopter and a pet owner, enabling them to create and manage pet listings. Adoption requests establish a relationship between users and pets, while interaction data is used to support the recommender system.

VI. Recommender System Design

Recommender systems are information filtering systems that predict user preferences and suggest relevant items. They are widely used in various domains such as e-commerce, healthcare, and online services [3].

In the EWAA system, a hybrid recommender approach is adopted to improve matching accuracy by combining collaborative filtering and content-based filtering techniques. This hybrid approach mitigates the limitations of individual methods by combining user preference modeling with interaction-based similarity.

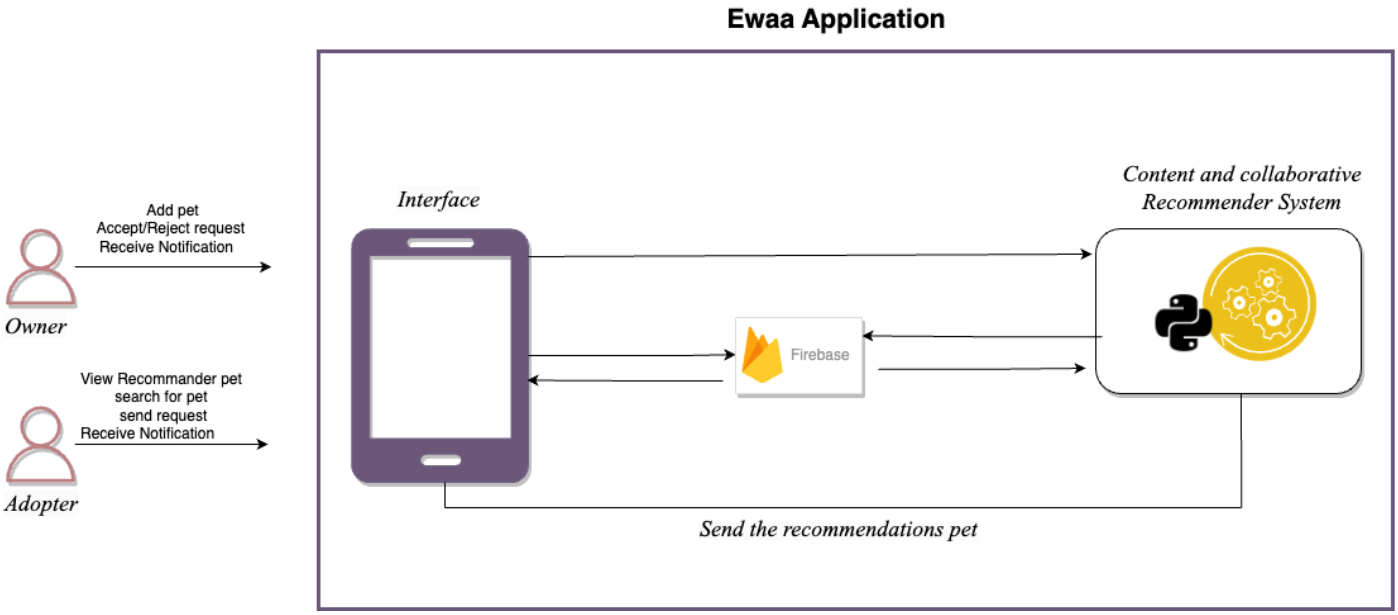


Fig. 2. EWAA system architecture

A. Collaborative Filtering

The collaborative filtering approach identifies users with similar preferences and recommends pets based on their interactions. The system constructs a user–item (likes) matrix R in which each row represents a user and each column represents a pet, with entries indicating whether the user has liked the pet. The similarity between the active user and every other user is computed using the cosine similarity measure:

$$\text{sim}(u, v) = \frac{R_u \cdot R_v}{\|R_u\| \, \|R_v\|} \tag{1}$$

A similarity threshold of $\tau = 0.5$ is applied to retain only sufficiently similar users, from which the top $K = 10$ neighbors are selected. The neighborhood size $K = 10$ and threshold $\tau = 0.5$ follow common practice in memory-based collaborative filtering, where small-to-moderate neighborhoods and mid-range similarity cut-offs are typically reported as a balance between recommendation accuracy and noise from weakly correlated users [3], [12]. Candidate pets already liked by the active user are excluded. The remaining candidates are scored by accumulating the similarity weights of neighbors who liked them, then normalized by the number of contributing neighbors. When the active user has no liked pets or no sufficiently similar users exist, the collaborative filtering component falls back to low-popularity or underexposed pets, to balance visibility and mitigate popularity bias. This choice prioritizes fairness and exposure over accuracy, which is suitable in adoption contexts. The procedure is summarized in Algorithm 1.

This algorithm recommends pets by identifying users with similar preferences and aggregating their interactions to generate relevant suggestions.

Algorithm 1 Collaborative Filtering Recommender

Require: Active user ID u, similarity threshold $\tau = 0.5$, neighborhood size $K = 10$
Ensure: Top 10 recommended pets
Load user–item (likes) matrix R
if u has no liked pets **then**
  return the 10 least-popular pets ▷ cold-start fallback
end if
for each user $v \neq u$ **do**
  $\text{sim}(u, v) \leftarrow \text{cosine}(R_u, R_v)$
end for
$S \leftarrow$ the top K users v with $\text{sim}(u, v) \geq \tau$
if S is empty **then**
  return the 10 least-popular pets ▷ no similar users
end if
Initialize $\text{score}[i] \leftarrow 0$, $\text{count}[i] \leftarrow 0$ for all pets i
for each user $v \in S$ **do**
  for each pet i liked by v and not by u **do**
    $\text{score}[i] \leftarrow \text{score}[i] + \text{sim}(u, v)$
    $\text{count}[i] \leftarrow \text{count}[i] + 1$
  end for
end for
for each pet i with $\text{count}[i] > 0$ **do**
  $\text{score}[i] \leftarrow \text{score}[i]/\text{count}[i]$ ▷ normalization
end for
Sort pets by score in descending order
return the top 10 pets

B. Content-Based Filtering

The content-based approach builds a user profile from the attributes of the pets the user has previously liked, and then recommends candidate pets whose attributes best match that profile. Four attribute groups are considered: category (cat, dog), breed, color, and age group. Each pet is encoded as a one-hot feature vector over the union of all attribute values [13],

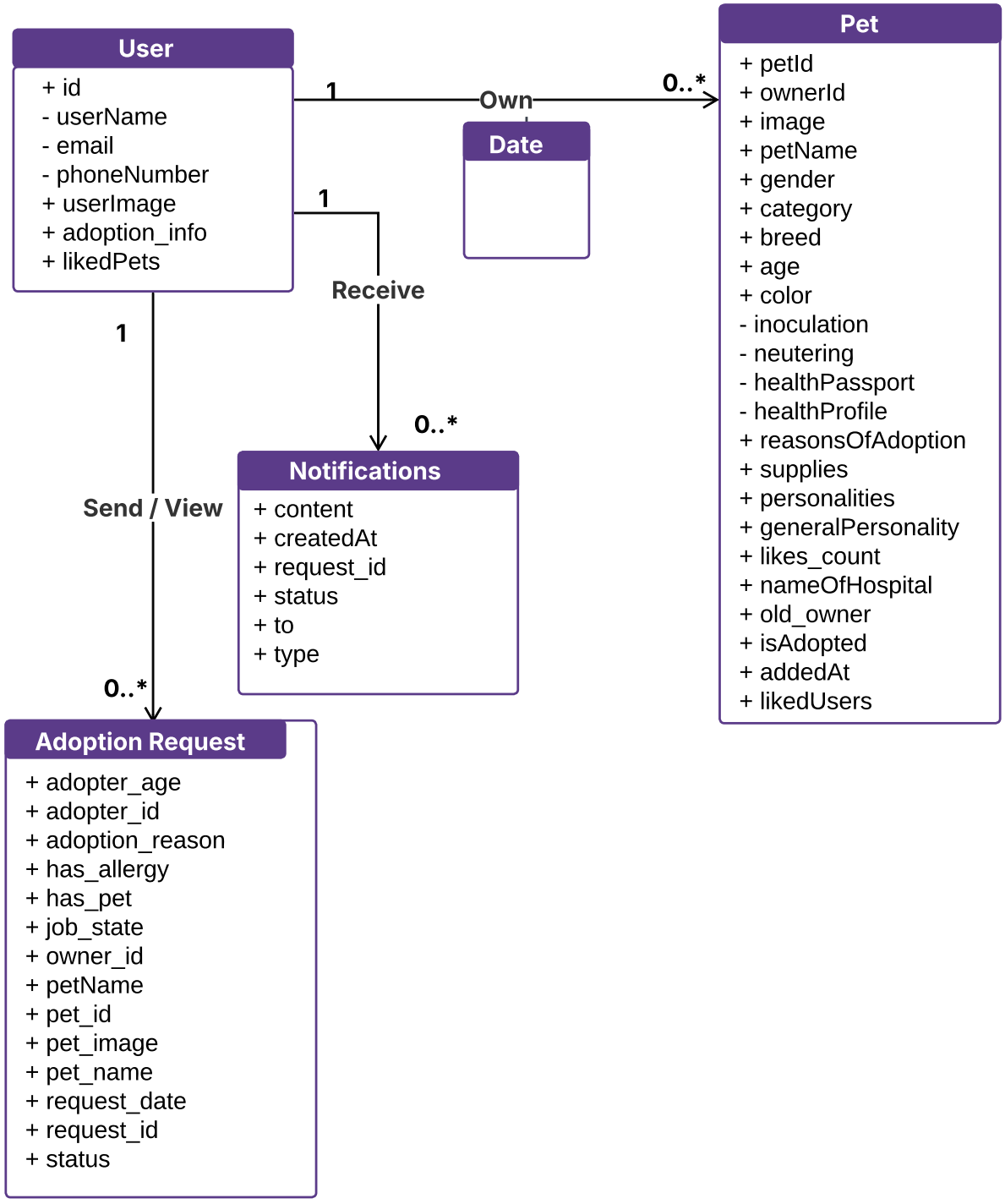


Fig. 3. Data model of the EWAA system.

forming a binary feature matrix F in which rows correspond to pets and columns to feature values.

The user profile p is constructed by averaging the one-hot vectors of the pets the user has liked:

$$p = \frac{1}{|L|} \sum_{\ell \in L} F[\ell] \tag{2}$$

where, L denotes the set of pets liked by the active user. This averaging yields frequency-based attribute weights for the user. For example, if a user has liked three pets with colors

Algorithm 2 Content-Based Recommender

Require: Active user ID u
Ensure: Top 10 recommended pets
$L \leftarrow$ pets liked by u
if L is empty **then**
  return the 10 most-popular pets ▷ cold-start fallback
end if
$F \leftarrow$ one-hot feature matrix over {category, breed, color, age}
Initialize user profile $p \leftarrow \mathbf{0}$
for each pet $\ell \in L$ **do**
  $p \leftarrow p + F[\ell]$ ▷ accumulate one-hot vectors
end for
$p \leftarrow p/|L|$ ▷ attribute-frequency weights
$C \leftarrow$ candidate pets (excluding L)
for each pet $c \in C$ **do**
  $\text{score}[c] \leftarrow \sum_k F[c,k] \cdot p[k]$
end for
Sort C by score in descending order
return the top 10 pets

$\{\text{white}, \text{white}, \text{brown}\}$, the profile assigns a weight of $2/3$ to white and $1/3$ to brown, while all other colors receive a weight of zero. The same principle applies to breed, age, and category, so that the profile captures the relative importance of each attribute value based on the user's observed preferences.

Each candidate pet c (excluding pets already liked by the user) is then scored as the weighted sum of its feature vector with the profile:

$$\text{score}(c) = \sum_k F[c,k] \cdot p[k] \tag{3}$$

Candidates are ranked by their scores and the top 10 pets are returned to the user. The full procedure is presented in Algorithm 2.

This algorithm recommends pets by matching candidate pet attributes to the user's preference profile derived from previously liked pets.

C. Hybrid Integration

The EWAA system implements a *mixed* hybrid configuration [2]: the content-based and collaborative filtering recommenders operate independently, and each produces its own top-10 list. Both lists are then presented to the user on the recommendation page of the application. This design was chosen for two reasons. First, it mitigates the cold-start limitation of collaborative filtering: new users with few or no liked pets still receive useful content-based recommendations. Second, it mitigates the over-specialization limitation of content-based filtering, since the collaborative component can surface pets that do not match the user's explicit attribute preferences but are enjoyed by users with similar overall taste. The two components are integrated through a Flask service that receives the active user's ID from the Flutter front end, executes both recommenders on the data retrieved from the Firebase database, and returns both recommendation lists to the application as JSON.

D. User Interface and Navigation

The EWAA application features a user-friendly interface designed to ensure ease of navigation and accessibility. The Home Page serves as the central hub, allowing users to browse pets, apply filters, and access key features such as adoption requests and user profiles.

Users can navigate through different sections including categories, pet details, notifications, and favorites. The system also supports account management features such as login, registration, and password recovery. Fig. 4 illustrates the application navigation flow.

E. Graphical User Interface

Fig. 5 presents sample screens from the EWAA application, including the login page, home page, filtering interface, and pet listing interface. The graphical user interface was designed to provide a simple and intuitive experience for both adopters and pet owners.

F. UX Guidelines

To ensure usability and accessibility, several user experience guidelines were considered in the design of the EWAA application:

- The system targets a high reliability rate, operating without failure in at least 95% of cases.
- The interface is designed to be intuitive and easy to navigate for users of different technical backgrounds.
- User privacy is protected for both adopters and pet owners.
- The system is scalable and capable of handling a growing number of users.
- Users can learn and use the system efficiently in a short time.

VII. System Testing and Evaluation

Since the EWAA system was not deployed at scale, real user interaction logs were not available for a large-scale offline evaluation of the recommendation algorithms. Instead, the overall system was evaluated using User Acceptance Testing (UAT) and Non-Functional Requirements (NFR) testing with real users, assessing usability, reliability, performance, and overall user satisfaction. Both qualitative and quantitative methods were employed to ensure a comprehensive assessment.

A. User Acceptance Testing

User Acceptance Testing was conducted using two complementary approaches: questionnaires and interviews. A total of 20 participants were involved in the questionnaire-based evaluation.

The demographic profile of the participants was analyzed to ensure a representative evaluation of the system. The majority of the participants were within the 18–25 age group (approximately 60%), followed by users aged 26 to 35 (30%) and a smaller proportion over 35 years (10%). In terms of gender

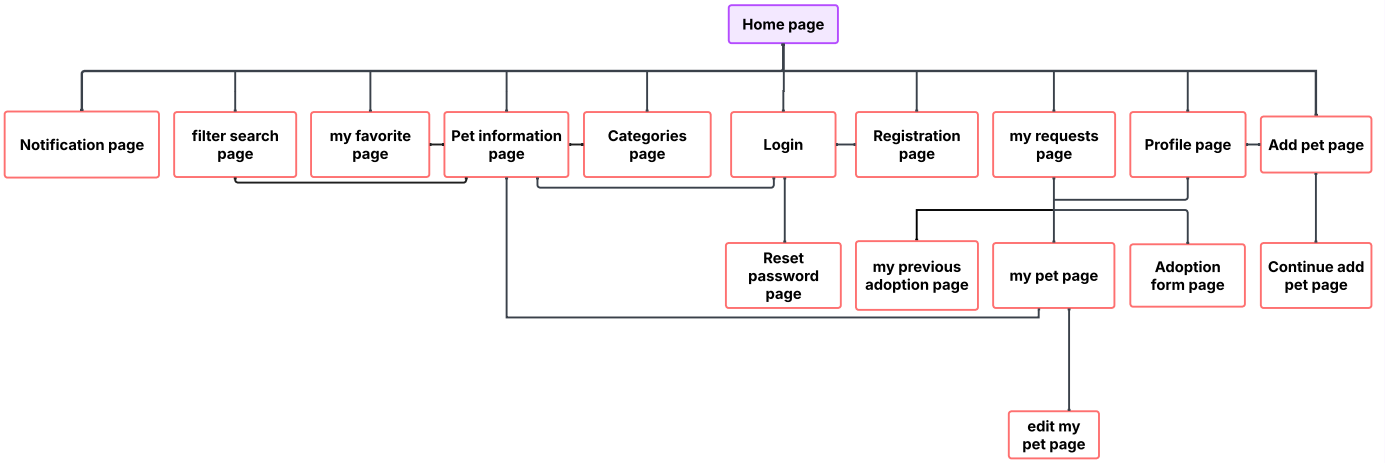
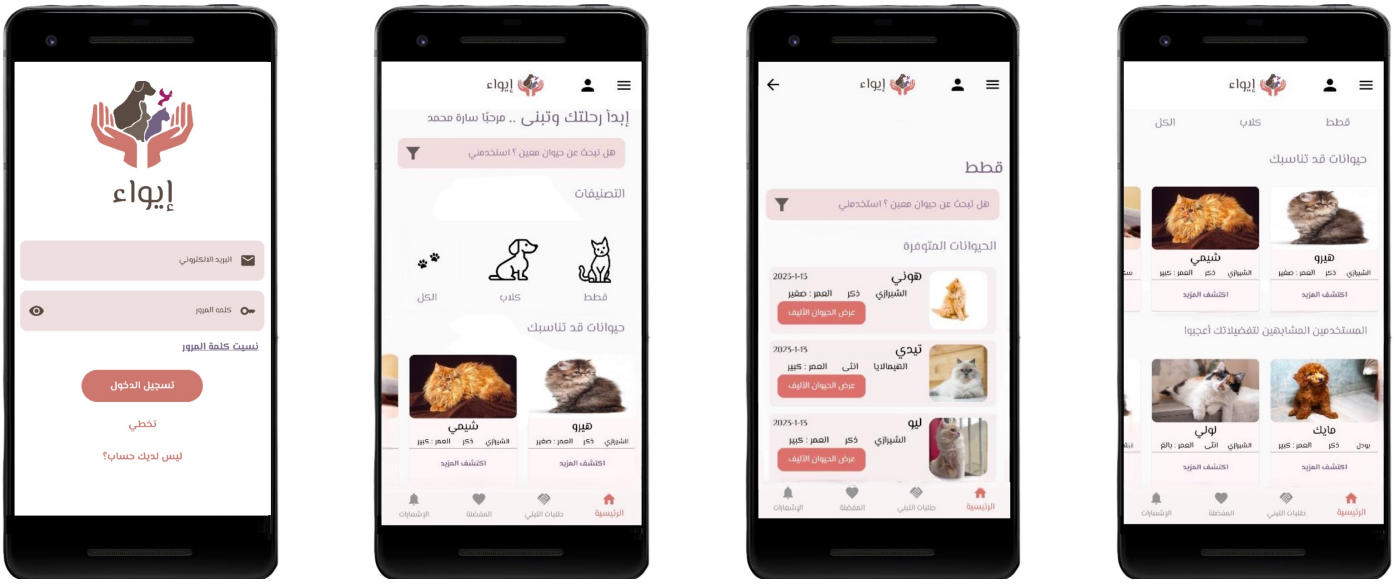


Fig. 4. Application site map



(a) (b) (c) (d)

Fig. 5. EWAA application interface

distribution, the sample included a balanced representation of male (45%) and female (55%) participants. Most participants (65%) reported having owned a pet previously, while 35% had no prior experience. This diversity ensured that the system was evaluated from both experienced and first-time adopter perspectives.

1) Questionnaire-based evaluation: Participants were asked to interact with the application and respond to a set of structured questions that evaluated usability, interface design, and system effectiveness.

The majority of the participants agreed that the application is easy to use, visually appealing, and effective in facilitating the pet adoption process. In addition, users confirmed that the system provides adequate information about pets and supports efficient communication between adopters and pet owners. The summarized results, presented in Table II, indicate a high level of user satisfaction with the EWAA application. Participants reported positive feedback regarding the interface design, usability, and overall functionality. In particular, users agreed that the system is clear, easy to use, and effective in facilitating the pet adoption process. Additionally, the recommender system was well received, with 85% of participants indicating that it helped them identify suitable pets based on their preferences. These findings demonstrate that the application meets user expectations and provides a positive user experience.

2) Interview-based scenario testing: To gain additional insights into system usability, 6 participants took part in interview-based testing. This testing was carried out using 11 predefined scenarios that captured the main functionalities of

TABLE II. SUMMARY OF QUESTIONNAIRE-BASED USER EVALUATION

Evaluation Statement	Agreement (%)	Interpretation
The application interface is attractive	100%	Very Positive
The application is easy to use	95%	Positive
The application is clear and understandable	95%	Positive
The system facilitates the adoption process	90%	Positive
Pet information is sufficient	90%	Positive
User privacy is maintained	100%	Very Positive
The recommender system is helpful	85%	Positive
Overall satisfaction with the application	100%	Very Positive

TABLE III. SCENARIO-BASED USER ACCEPTANCE TESTING RESULTS

No.	Test Scenario	Successful Users (out of 6)	Status
1	User registration	6/6	Pass
2	Add pet for adoption	6/6	Pass
3	Browse pets	6/6	Pass
4	Filter search results	6/6	Pass
5	View pet details	6/6	Pass
6	Add to favorites	6/6	Pass
7	Submit adoption request	6/6	Pass
8	Manage adoption requests	6/6	Pass
9	View notifications	6/6	Pass
10	View recommendations	6/6	Pass
11	View adoption history	3/6	Partial Pass

the application. The scenarios covered tasks such as registering a user, adding pets, filtering search results, handling adoption requests, and using the recommender system.

The results in Table III show that the system achieved a 100% task completion rate for 10 out of 11 test scenarios, indicating a high level of usability and functional correctness. The only scenario that presented difficulty was "viewing the adoption history", where some users were unable to locate the feature due to unclear navigation. This indicates a minor usability issue related to interface design rather than system functionality.

Overall, participants were able to complete core tasks such as signing up, adding pets, filtering searches, sending requests, and interacting with recommendations without any significant issues, demonstrating the system's usability and functional correctness.

B. Non-Functional Requirements Testing

Non-functional testing was conducted to evaluate key quality attributes of the system, including performance, learnability, and reliability. The results are summarized in Table IV.

1) Performance: System performance was evaluated by measuring page loading times during user interaction. All application pages loaded within a range of 1 to 12 seconds, satisfying the predefined requirement of a maximum loading time of 20 seconds. This indicates that the system provides a responsive user experience under the tested conditions.

2) Learnability: Learnability was assessed by measuring the time required for users to understand and complete core functionalities. Participants were able to learn and interact with the system within 1 to 8 minutes, which is well within the acceptable threshold of 10 minutes, suggesting that the system is intuitive for new users.

TABLE IV. NON-FUNCTIONAL REQUIREMENTS TESTING RESULTS

Quality Attribute	Measure	Target	Result
Performance	Page load time	$\leq$ 20 sec	1–12 sec
Learnability	Learning time	$\leq$ 10 min	1–8 min
Reliability	Task success rate	$\geq$ 95%	100%

3) Reliability: Reliability was evaluated by measuring the proportion of functionalities that executed without critical errors. Across the 11 interview-based scenarios, all core functionalities executed without runtime failure; the only observed issue was a navigation visibility problem in one scenario, which is a user-interface concern rather than a functional failure. On this basis, the system met the predefined reliability target of at least 95% task success at the functional level.

C. Discussion

The evaluation results, summarized in Table IV, indicate that the EWAA system meets its usability, performance, and reliability targets. The integration of core adoption functionalities with a personalized recommendation component supports the pet adoption workflow in an Arabic-speaking context.

The only identified limitation in the user-facing evaluation was related to the visibility of the adoption history feature, which can be addressed through minor interface improvements such as clearer navigation or dedicated menu options.

Several limitations of the evaluation should also be acknowledged. First, the sample size (20 questionnaire participants and 6 interview participants) is modest and drawn primarily from a single demographic group (60% aged 18–25), which limits the generalizability of the usability findings to the broader population of potential adopters. Second, participants interacted with the system in controlled testing scenarios rather than through natural long-term use, which may lead to overestimation of reported satisfaction. Third, the questionnaire-based dataset used to bootstrap the recommender reflects owner-reported pet preferences rather than real adoption behavior, which may introduce bias into the learned user profiles once the system is deployed. Finally, the current evaluation focuses on system usability and non-functional qualities rather than the offline accuracy of the recommendation algorithms themselves; the 85% perceived helpfulness reported in Table II is a subjective measure and does not quantify recommendation quality.

A dedicated quantitative evaluation of recommendation quality, using metrics such as precision@N or hit rate on held-out interaction data, is left as future work once sufficient real-user interaction data are collected post-deployment. Overall, the findings indicate that EWAA provides a user-centered platform that supports the pet adoption process, and that addressing the identified navigation issue will further improve the user experience.

VIII. CONCLUSION

This study presented EWAA, an Arabic-enabled mobile platform for pet adoption in Saudi Arabia that integrates a hybrid recommender system combining content-based and collaborative filtering. The content-based component builds a user

profile from one-hot encoded pet attributes of previously liked pets, while the collaborative filtering component identifies similar users through cosine similarity and recommends pets based on their preferences. The two components operate in a mixed hybrid configuration, which mitigates the cold-start limitation of collaborative filtering and the over-specialization limitation of content-based filtering.

The main scientific contribution of this work is the integration of a mixed hybrid recommender, combining one-hot attribute-based content filtering with cosine-similarity collaborative filtering, into an Arabic-language mobile adoption platform, addressing a gap in prior recommender approaches that have neither combined complementary signals within deployed adoption systems nor targeted Arabic-speaking users.

The evaluation results indicate that the system meets its usability, performance, and reliability targets. User Acceptance Testing with 20 questionnaire participants and 6 interview participants confirmed 100% task completion on 10 of 11 test scenarios, with the remaining scenario identifying a minor navigation issue. Non-Functional Requirements testing further verified that page load times (1–12 seconds) and learning times (1–8 minutes) fell well within their predefined targets.

Future work will include an offline evaluation of the recommender component using standard metrics such as hit rate and precision once sufficient real-user interaction data are collected post-deployment. Further directions include exploring alternative recommendation models such as item-based collaborative filtering or matrix factorization, extending the feature set beyond category, breed, color, and age, and adding diversity-aware post-processing. Planned platform extensions include pet health tracking and adoption analytics.

ACKNOWLEDGMENT

The author thanks Rawan Alqarni, Hadeel Quayyid, Gadah Alotibi, and Rahaf Almoqbel for their work on the implementation and testing of the EWAA system.

DECLARATION ON GENERATIVE AI

During the preparation of this manuscript, the author used AI-based language editing tools to assist with grammar, clarity, and formatting. The author reviewed and edited all AI-assisted content and takes full responsibility for the originality, accuracy, and integrity of the final manuscript.

REFERENCES

[1] K. J. Scoresby, E. B. Strand, Z. Ng, K. C. Brown, C. R. Stilz, K. Strobel, C. S. Barroso, and M. Souza, "Pet ownership and quality of life: A systematic review of the literature," *Veterinary Sciences*, vol. 8, no. 12, p. 332, 2021.

[2] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[3] C. C. Aggarwal *et al.*, *Recommender systems*. Springer, 2016, vol. 1.

[4] R. Karani, P. Tawde, N. Popovich, J. Patel, S. Doshi, and R. Mansuria, "Petpaws: A comprehensive dataset and recommender system for canine and feline breeds," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*. IEEE, 2023, pp. 1–6.

[5] R. G. de Luna, E. M. Latayan, A. V. O. Cuizon, M. J. M. Onte, J. P. D. Umali, and H. M. Ortega, "Pawfect match: A user-interactive dog breed recommendation powered by machine learning," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2024, pp. 97–102.

[6] M. R. Q. Montiveles, E. J. M. Fernandez, B. H. C. Duron, and C. P. Barbosa, "Pawster Care: An Android-based pet cruelty reporting and adoption application with recommender system," *UM Research Journal*, 2021.

[7] M. Chaudhari, S. Parab, S. Ghute, R. Chaudhari, and K. Deshpande, "Enhancing pet care through petree: A machine learning based comprehensive framework," in *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, 2024, pp. 339–351.

[8] E. Afoura, "PetMatch: Pet adoption (mobile application)," Google Play Store, 2026, android mobile application; last updated 5 April 2026; accessed 2025-02-19. [Online]. Available: https://play.google.com/store/apps/details?id=com.metron.petmatch

[9] Nestlé, "Petfinder: Adopt a pet (mobile application)," Google Play Store and Apple App Store, 2025, android and iOS mobile application; accessed 2025-02-19. [Online]. Available: https://play.google.com/store/apps/details?id=com.discovery.petfinder

[10] Market Push Apps SRL, "Pet adoption (wix web application)," Wix App Market, 2025, web-based application for the Wix platform; accessed 2025-02-19. [Online]. Available: https://www.wix.com/app-market/web-solution/pet-adoption

[11] A. Srivastava, S. Bhardwaj, and S. Saraswat, "Scrum model for agile methodology," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 864–869.

[12] J. L. Herlocker, J. A. Konstan, and J. Riedl, "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms," *Information Retrieval*, vol. 5, no. 4, pp. 287–310, 2002.

[13] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer, 2007, vol. 4321, pp. 325–341.

State-AttentNet: A Dynamic Volatility-Adaptive Hybrid Framework for Market State Classification in Frontier Economies

Md. Abul Kalam Azad[1], Abdul Kadar Muhammad Masum[2]*, Najmus Saadat[3],
Esrat Jahan[4], Ramona Birău[5], Virgil Popescu[6], Iuliana Carmen Bărbăcioru[7], Ștefan Mărgăritescu[8]
Dept. of Business and Technology Management, Islamic University of Technology, Gazipur, 1704, Bangladesh[1]
Dept. of Computer Science and Engineering, Southeast University, Dhaka, 1208, Bangladesh[2]
Computational Intelligence Lab, Southeast University, Dhaka, 1207, Bangladesh[3,4]
"Eugeniu Carada" Doctoral School of Economic Sciences, University of Craiova, Craiova, Romania[5,8]
Faculty of Economic Science, Constantin Brâncusi University of Târgu Jiu, 210185 Târgu Jiu, Romania[5]
Faculty of Economics and Business Administration, University of Craiova, Craiova, Romania[6]
Faculty of Engineering, University of Târgu Jiu, Gorj County, Romania[7]

***Abstract*—The proper identification of latent market states is important for effective risk management. However, existing frameworks are often not good at distinguishing the difference between stable situations and systemic crashes in the high entropy environments of frontier economies. To address this challenge, this study presents State-AttentNet, an application-driven market-state classification framework for the Dhaka Stock Exchange. The framework combines a volatility-adaptive labeling scheme with a bidirectional LSTM encoder and temporal attention to classify stable and crash-like market states in the Dhaka Stock Exchange. The market sensitive technical features are synthesized with Bidirectional LSTM encoder of the temporal context into a hierarchical pipeline proposed methodology. This model is also supplemented by Adaptive Temporal Attention model to bring focus to high impact volatility events in a rolling window. Here, 'crash' means an adaptive volatility-stress state, not a return direction. Empirical evaluation on a longitudinal 26-year dataset shows that the model achieves 93% classification accuracy and an AUC of 0.97. It outperforms traditional baseline models in discriminating between crash and stable states. The results have significant practical implications for institutional investors as it is a trusted, automated detection instrument. This system is conducive for state contingent risk reduction measures and minimizing false alarms in instances of correction of incidental markets. Lastly, SHAP used to check the operational integrity of the model. This suggests that structurally informative lag signals and not stochastic noise determine classification decisions.**

Keywords*—Frontier markets; hybrid deep learning; adaptive attention mechanism; dynamic volatility labeling; Explainable AI (XAI)

I. Introduction

Financial markets are rarely in a steady state. Instead, they evolve through regimes in which expected returns, volatility and cross asset dependence change between calm, transitional and stress periods. This view is often formalized by latent states whose parameters change according to a Markov process, which makes it possible to model abrupt but persistent changes in a tractable way [1]. These regime shifts are directly relevant for risk, because correlations and volatilities are likely to increase in adverse states, which affects the diversification benefit and the cost of holding risky exposures over time [2]. Recent empirical evidence also suggests that regime detection is depends on more than just price fluctuations. It requires signals that capture the re-organization of the joint structure of the market under stress. For example, the information about realized covariance can be useful to enhance the regime identification when the standard models face difficulties, which implies that dependence dynamics should be considered as state variables in risk monitoring [3]. Regime switching specifications also enhance the realized volatility forecasting since parameters can adjust as the markets switch across volatility states [4]. At a broader macro finance scale, probabilistic models have been utilized to test the validity of bear markets and recessions occurring within a country simultaneously, reinforcing the view that regime events can have international transmissions but can largely not occur at the same time [5]. Together these studies make market state classification a fundamental task in understanding the risk in the case of non-stationarity rather than a purely descriptive task.

Building on these foundations, modern machine learning provides flexible tools for modelling nonlinear and time varying market conditions. Multi Transformer architectures have been proposed in volatility forecasting and risk assessment and have shown great potential to learn complex temporal interactions in unstable periods [6]. Hybrid machine learning designs are also highlighted for volatility prediction for risk management because the combination of complementary learners can offer superior performance compared to single learners in the context of a noisy and structurally changing environment [7]. Emerging hidden state formulations such as the Hidden Quantum Markov Model imply that richer state representations may capture aspects of regime dynamics that classical structures may not be able to express [8]. Related work also indicates that adaptive Transformer reinforcement learning frameworks can interpret sentiment-driven signals into predictive and decision oriented policies that respond to

*Corresponding author

changing market conditions [9].

In parallel, the feature space of market modeling is growing towards multimodal and explainable pipelines which are closer to the real decision setting. Temporal feature attention along with GraphSAGE on multimodal financial data shows that attention and graph encoders can learn predictive structure from interrelated sources of information [10]. Recent studies highlight explainability-guided recurrent modeling using SHAP values, demonstrating that strong performance should be accompanied by transparent reasoning under regime shifts [11]. Topology-driven CNN designs that fuse news sentiment with market dynamics are similarly suggestive of the ability of information flow and network effects to materially shape predictive behavior [12]. In high frequency situations, generative discriminative learning based on hidden state structure and discriminative boundaries has been shown to be useful in enhancing the regime separation in the situation of microstructure noise, demonstrating the importance of hybrid probabilistic and machine learning pipelines for the classification of states [13].

Despite these improvements, there remains an important gap for frontier equity markets over long horizons. Much of the recent evidence is developed market-focused, based on high frequency data sets, or assesses performance over short windows though frontier markets suffer from structural breaks and changing liquidity and participation driven shocks that may distort fixed labelling schemes and reduce the transferability of models. At the same time, theory-driven portfolio research focuses on how regimes are usually unobservable and have to be inferred with regards to filtered probabilities in order to support dynamic decision rules, which makes classification explicitly decision relevant rather than only predictive [14]. This motivates regime classifiers that can adapt labels to evolving risk and be robust under long run non stationarity and still provide interpretability for monitoring.

To overcome these limitations of the previous studies, the goal of this study is not to predict prices, but to explicitly classify the market states in Dhaka Stock Exchange using the daily data between 1999 and 2025 [15]. This study propose a dynamic volatility threshold labeling approach to the definition of stable and crash states that is adaptive to time varying market risk. The proposed framework incorporates market-sensitive technical features and propose State-AttentNet, which is a bidirectional recurrent sequence model with attention that focuses on the most informative days within each rolling window instead of all lag equally. Finally, Explainable AI (XAI) techniques are employed to ensure that the decisions of the model are not based on transient noise, but on meaningful lag signals. Experimental results demonstrate that State-AttentNet yields better classification performance compared to common recurrent baselines underscoring the use of State-AttentNet for transparent regime monitoring in a frontier market setting for risky regime aware decision making. The contribution of this study is not to introduce a fundamentally new neural architecture, but to develop an application-driven framework that integrates adaptive volatility labeling, temporal attention and explainability for long-horizon market state classification in a frontier market context.

II. Literature Review

The development of financial markets is rarely in a smooth and continuous way. Rather, emotional changes change with time, with periods of calmness punctuated by spurts of stress, volatility peaks, and dependence reversals. This set of recurring transitions has been typically described as market regimes, in which the properties of returns and risk are restructured across a set of discrete states as opposed to being due to a unique stable process. Particularly, risk management of a regime has two extensions. Regime-switching volatility models combine ARCH-type volatility clustering with latent regime changes, which enable volatility persistence to vary between regimes [16]. Less specific regime switching specifications allow the conditional distribution to differ under a Markov state process, thus allowing the mean, dispersion, and tail behavior to vary across states [17]. The volatility evidence of recent regime switching also suggests that volatility dynamics can also be regime-specific to other designs of switching between regimes that use alternative score driving designs, meaning that risk boundaries are not to be fixed as long as structural conditions change [18]. Combined, these contributions explain why the single-regime assumption may be structurally misspecified when volatility and tail risk rearrange during stress.

A first cluster of studies deals with how regime should be defined and segmented in a way that is robust, transparent and reproducible. Zegadło *et al.* [19] suggests a rule-based bull-bear dating procedure with minimum tuning, which favors label stability and the comparability between studies. Wang *et al.* [20] goes even further by adding a four-state structure through the HSMM model, demonstrating that markets may survive in intermediate conditions and that state duration characteristics enhance the validity of segmentation. Maheu *et al.* [21] similarly motivates multi-state design but focuses on real-time dating and forecasting utility as showing that regime forecasts will be particularly useful in windows of crisis since timing risk exposure is when it will matter most. Collectively these works support multi-state frameworks, however there is one limit that continues to prevail. State definitions and boundary choices are often tied to particular modelling constraints or stylization of mature market behaviour, leaving the transferability to structurally unstable markets open. A second strand reinforces regime identification by introducing forward-looking information that may adjust ahead of spot returns. Lai *et al.* [22] shows that option implied horizon spread information outperforms returns-only inputs for regime detection, supporting the concern that derivative markets contain anticipatory information on risk. However, the approach is data-intensive and assumes deep and liquid options markets. Such infrastructure is not necessarily available in frontier exchanges, diminishing feasibility at the very places where regime monitoring in a timely manner may be most important.

A third body combines probabilistic regime structure with deep learning, where state probabilities form informative inputs or scaffolding of flexible sequence models. Another regime that Erlwein-Sayer *et al.* [23] offers to deep learning template is HMM state probability extraction followed by input into an LSTM to improve his results in terms of forecasting European corporate credit spreads. Haase and Neuenkirch [24] deal with instability by combining forecasts with regime forecasts, indicating that regime forecasts enhance

predictability in bull-bear regimes and returns in the US, which diminishes the vulnerability to regime specific model error. An interpretable HMM-based market situation estimator coupled with a trading strategy is created by Chen *et al.* [25] and is a useful benchmark of regime classification despite its feature design being ahead of modern representation learning. Although people demonstrate that hybridization is fruitful, the empirical evidence is largely developed market-oriented and tends to focus more on forecasting aim than on explicit daily classification of long-horizon under the changing volatility baselines.

Fourth cluster is an extension of regime forecasting to multimodal and text informed systems, indicating that market states are also determined by information diffusion and narrative shocks along with the price movement. Battazza *et al.* [26] suggests a complete end-to-end market state prediction pipeline comprising of ontology-based asset selection, non-stationary Markov chain, BERT sentiment representations, and LSTM modeling, as well as it is shown how regime inference can be incorporated into a larger allocation mechanism. The same author Mudarisov *et al.* [27] also states that LLM or FinBERT style news augmentation is better at regime prediction compared to time series pure baselines across industries, aligning with broader machine learning investigations that demonstrate the substantial impact of news sentiment on market dynamics [35]. Though promising, these methods are structurally weak in frontier settings when historical records might be biased, language diffusion might be uneven, as well as heterogeneity in news coverage can induce bias in learnt regime cues.

Recent developments have moved away the focus of prediction-focused testing to decision-focused assessment that is more in line with the trends in algorithmic trading and portfolio building in developing contexts. This fact is supported by Hanauer and Kalsbach [28] that machine learning models are capable of finding economically significant signals in a wide emerging-market equity space, which in turn makes the argument that non-linear predictability can be translated to implementable strategy value within realistic constraints. Studies that are strategy-focused also review machine learning as an engine of signals of algorithmic investment rules on both global and Central-Eastern European equity indices where technical characteristics are mapped into systematic decision regulations and measured within a consistent performance procedure [29]. In addition to equity signals at the daily-frequency, short-horizon studies of the Asian markets show how the data-driven models may be implemented into the rule-based pipelines of entry and exit decisions. The use of deep learning architectures in high-frequency pairs trading studies in Taiwan to predict the boundaries of spreads and issue trading clues supports the increased preference to use decision-relevant validation over accuracy-only prediction [30]. Collectively, these results give stronger empirical impetus to regime-sensitive targets, since algorithmic strategies are directly sensitive to volatility changes and to changes in the structure and tenacity of market regimes.

The other direction conceptualizes regimes as labels, as well as state dependent mechanisms, which change the drivers that are economically relevant. Pakštaitė *et al.* [31] demonstrates that macroeconomic influences of the Bitcoin differ across regimes with Bayesian selection HMM backbone which confirms the general suggestion that the relevance of factors can change among states. This is furthered by Wang *et al.* [32] who incorporates heteroskedastic networks with HMM and machine learning to identify regime switch and early signals, which suggests that transitions can be encoded in evolving interaction structure, as opposed to volatility. These papers reinforce the arguments in favor of adaptive, state contingent modelling but validation is still focused on individual large markets or crypto-environment, where frontier equity generalization has not been sufficiently evaluated. Shu *et al.* [33] comes up with asset specific regime labels based on jump model logic, regime forecasts based on GBDT and presents better results of dynamic allocation, which points to the importance of regime forecasts to enter portfolio policy. The use of an ensemble pipeline of machine learning and HMM with a voting classifier is suggested by Gupta *et al.* [34] and has a higher regime shift detection that is more stable and could be used to aid trading decisions in ETF-based contexts. These allocation-based evaluations are however normally carried out in developed market universes or proxy assets. They seldom treat long-horizon daily regime classification in which volatility baselines, liquidity situations, and microstructure of the market change significantly over time.

Across these studies, four gaps recur. First, long-horizon daily regime classification in the frontier equity markets is still limited, as most of the evidence is pegged to the United States or Europe, in the ETF proxies, in the options based identification or in the news rich settings. Second, there is rigidity in labelling, making the boundaries of the state in most cases fail to explicitly adjust with changes in volatility baselines that develop slowly over decades. Third, not all hybrid models are always associated with interpretability even though the regime signals are applied in risk management context where transparency is valued. Fourth, the feasibility of the data limits transfer as frontier markets seem to have only dependencies and news intensive regimes, which can be infeasible in long gaps of history.

To address these limitations, this study focuses on application-driven market-state classification in a frontier-market setting. Using 26 years of daily Dhaka Stock Exchange data [15], this study develop a causally labeled regime-classification pipeline based on a dynamic volatility-adaptive thresholding scheme. A bidirectional recurrent model with temporal attention is then applied to identify informative periods within each rolling window. Finally, this study incorporate explainability analysis to improve transparency and support decision-relevant interpretation in a data-constrained frontier-market environment.

III. Methodology

This study proposes an end-to-end deep learning architecture to classify the stable and crash market regime. As illustrated in Fig. 1, the general methodology includes data preparation, feature extraction, sequential modeling based on Bi-LSTM encoder and temporal attention mechanism and a classification head to reliably predict the final market state.

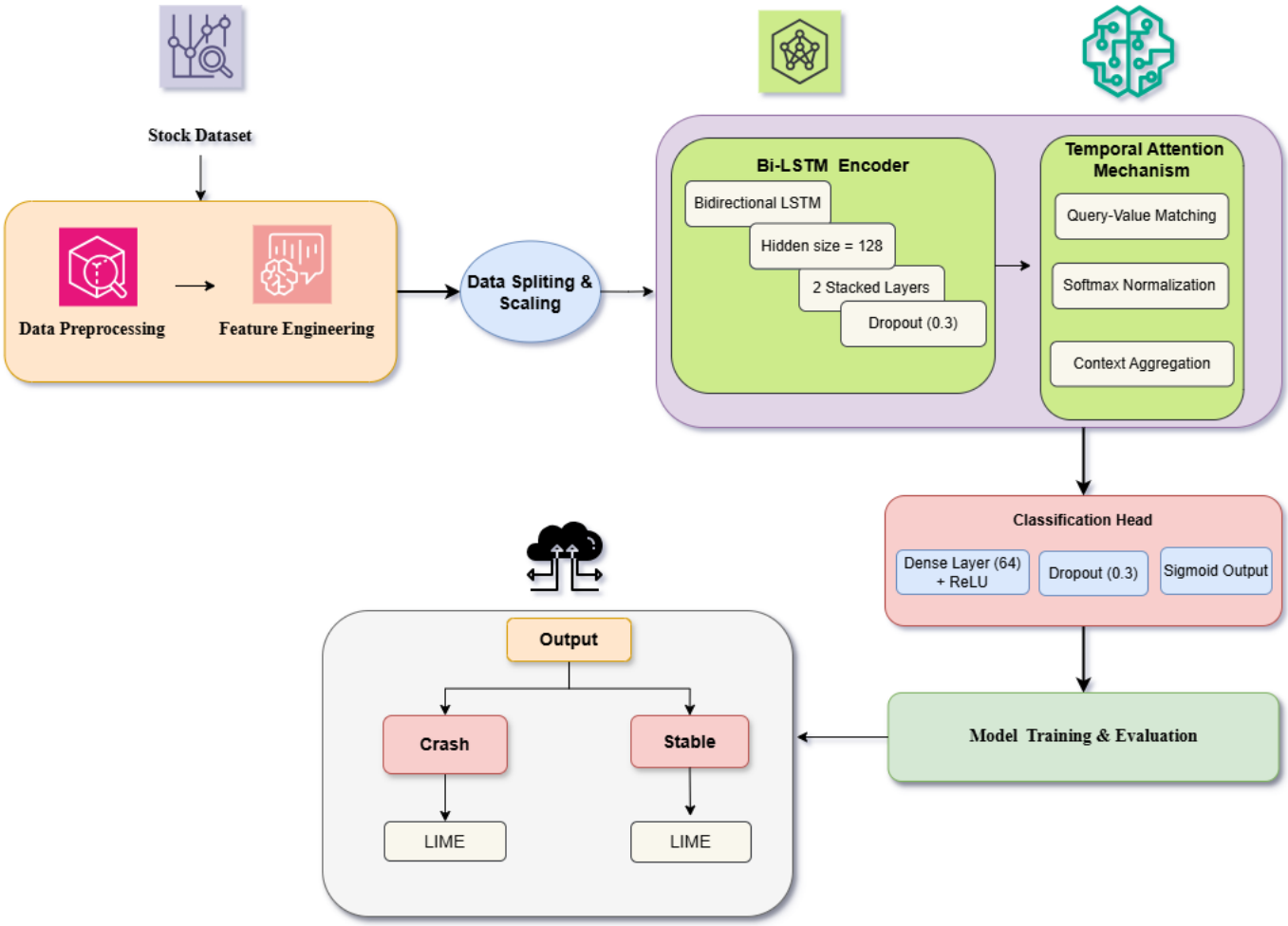


Fig. 1. End-to-end workflow of the State-AttentNet architecture.

A. Dataset Description

This study utilizes a high-fidelity dataset from the Dhaka Stock Exchange (DSE) spanning a 26-year period from 1999 to 2025. The repository contains over 1.5 million trading records across 534 distinct securities. As illustrated in Fig. 2, the market's price history exhibits extreme instability and frequent structural breaks, such as the 2010 capital market crash and the 2020 pandemic-induced volatility. Unlike developed markets, the DSE is characterized by high-entropy price movements and significant volatility clustering. These complex dynamics and heavy-tailed returns provide a rigorous benchmark for evaluating this proposed framework's ability to distinguish between systemic shocks and incidental market noise.

To empirically quantify these regime shifts, Table I presents a chronological breakdown of the distinct phases of the market that were seen within the dataset. This tabulation describes the development of volatility from the contrasting periods of systemic distress like the 2011 crash to regulatory interventions like the floor price era, thereby validating the non-stationary nature of the market. Specifically, the table calls attention to the fact that the market risk changes drastically through time. For example, the average volatility at the time of the 2011 crash was extremely high at 0.344, while in the period from 2022 to 2023 in the regulation era, volatility fell to only 0.035. This huge gap proves that the market does not act the same way every year. Because of these wide variations, a fixed-fire standard model with fixed rules would not predict the market correctly. This evidence justifies why we need a dynamic model that can automatically adapt to these changing conditions whether the market is in a bubble, a crash or a stable phase.

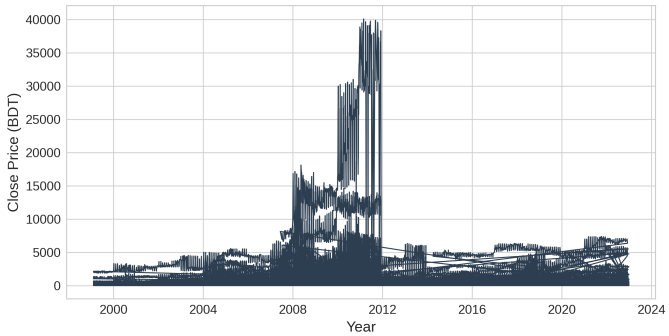


Fig. 2. Long-term price history of the DSE dataset.

TABLE I. CHRONOLOGICAL EVOLUTION OF MARKET STATES AND VOLATILITY DYNAMICS (1999–2024)

Year	Market Phase	Avg. Volatility	High-Risk Days	Key Event
1999–2000	Early Development	0.097	20.9%	Initial market formation volatility.
2004	Pre-Crisis Bubble	0.155	26.1%	Early signs of market overheating.
2010	The Great Bubble	0.215	24.1%	Highest speculation period.
2011	The Crash	0.344	25.9%	Record volatility due to market collapse.
2012–2016	Stabilization	0.118	22.5%	Post-crash recovery and consolidation.
2018	Election Effect	0.102	27.1%	Uncertainty due to national elections.
2020	Pandemic Shock	0.123	25.1%	Global volatility due to COVID-19.
2022–2023	Regulation Era	0.035	16.3%	Floor Price artificially reduced volatility.
2024	Correction Phase	0.025	26.9%	Removal of floor price triggered volatility.

B. Data Preprocessing and Adaptive State Identification

The transformation of raw DSE sequences into a structured dataset involves a systematic pipeline to ensure numerical stability and contextual relevance. To maintain chronological continuity without look-ahead bias, missing values and zero-entry anomalies are addressed using the Forward Fill propagation method. Given the high-entropy nature of emerging markets, Robust Scaling is employed to normalize the feature space. This technique utilizes the median and Interquartile Range, effectively mitigating the impact of extreme price spikes while preserving significant market signals. The operation is defined, as in Eq. (1):

$$x' = \frac{x - median(x)}{IQR(x)} \tag{1}$$

The frontier markets like the DSE are usually characterized by liquidity gaps and discontinuous trading that may cause artificial breaks in observed returns. Such markets also exhibit fat-tailed returns thus Robust Scaling is the choice of scaling which would eliminate the effect of extreme shocks. A notable aspect of this framework is the departure from traditional static crash rules in favor of a Dynamic Rolling Threshold for state labeling. Let P_t denote the daily closing price and define "the log return", as in Eq. (2):

$$r_t = \ln(P_t) - \ln(P_{t-1}) \tag{2}$$

The 20-day rolling volatility is computed as the sample standard deviation of returns over the most recent 20 trading days, as given in Eq. (3):

$$\sigma_t^{(20)} = \sqrt{\frac{1}{19}\sum_{i=0}^{19}(r_{t-i} - \overline{r_t^{(20)}})^2}, \quad \overline{r_t^{(20)}} = \frac{1}{20}\sum_{i=0}^{19} r_{t-i} \tag{3}$$

In this study, Crash refers to a volatility-spike regime rather than a return-sign definition, and, therefore, it does not necessarily imply negative returns. To adapt the crash boundary to the local volatility environment, a time-varying threshold is derived from the trailing 60-day empirical distribution of volatility. Let $Q_{0.80}(\cdot)$ denote the empirical 80th percentile. The dynamic threshold is defined, as in Eq. (4):

$$\tau_t = Q_{0.80}\left(\{\sigma_{t-j}^{(20)}\}_{j=0}^{59}\right) \tag{4}$$

The choice of the 80% percentile threshold was empirically validated to balance sensitivity to volatility spikes and robustness to noise. A sensitivity analysis was conducted using alternative thresholds (75%, 80%, and 85%), where 75% resulted in excessive false positive crash signals, while 85% reduced the model's ability to detect meaningful volatility stress events. The 80% threshold provided the most stable and balanced performance. Market regimes are then labeled by comparing the current volatility against this adaptive threshold, as in Eq. (5):

$$\text{State}_t = \begin{cases} 1 \text{ (Crash)}, & \text{if } \sigma_t^{(20)} > \tau_t \\ 0 \text{ (Stable)}, & \text{otherwise} \end{cases} \tag{5}$$

After sorting the data by Trading_Code and Date, zero entries in Open, High, Low, Close, and Volume were treated as missing and forward-filled within each trading code to preserve temporal continuity. Observations lacking sufficient history to compute the 20-day rolling volatility, the 60-day volatility quantile used for dynamic labeling, the 14-day RSI, or the 60-day input sequence were excluded. This explains the apparent exclusion of early-period observations (e.g., 1999–2009), which do not satisfy the minimum history requirements for rolling statistics and sequence construction.

As illustrated in Fig. 3, this adaptive mechanism enables the architecture to contextually distinguish systemic structural breaks from incidental stochastic noise, as the threshold co-moves with the prevailing volatility baseline rather than remaining fixed across heterogeneous market phases. For reproducibility, all rolling statistics and quantiles are computed exclusively using trailing windows within each trading code, and observations that cannot form the required rolling windows for $\sigma_t^{(20)}$ and τ_t are excluded from the final labeled sample.

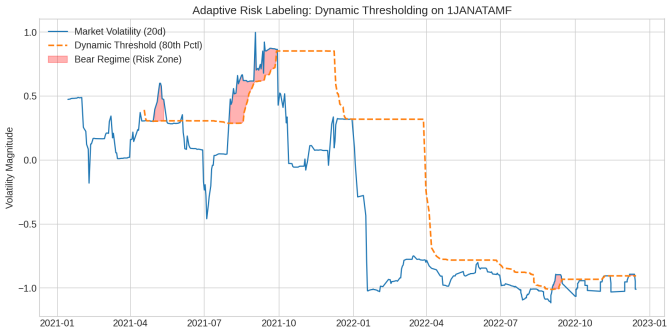


Fig. 3. Adaptive risk labeling logic using dynamic volatility thresholding.

C. Advanced Feature Engineering

To enhance the discriminative power of the proposed architecture, a multivariate feature set is constructed by capturing the temporal dynamics and liquidity shocks of the DSE. Instead of relying on raw price levels, the input vector is expanded to include indicators of momentum, volatility, and market pressure. The primary component involves the calculation of Logarithmic Returns to ensure mean-stationarity, supplemented by the Relative Strength Index and High-Low Spread to represent price dispersion. The statistical integrity and independence of this engineered feature set are validated through a correlation analysis. As shown in Fig. 4, the heatmap confirms that the selected attributes provide non-redundant information, which is essential for the model to distinguish between structural shocks and incidental market noise. Furthermore, to quantify the impact of trading activity on price movements, a composite Volume Force feature is incorporated, defined as in Eq. (6):

$$\text{Vol_Force}_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \times \ln(\text{Volume}_t + 1) \tag{6}$$

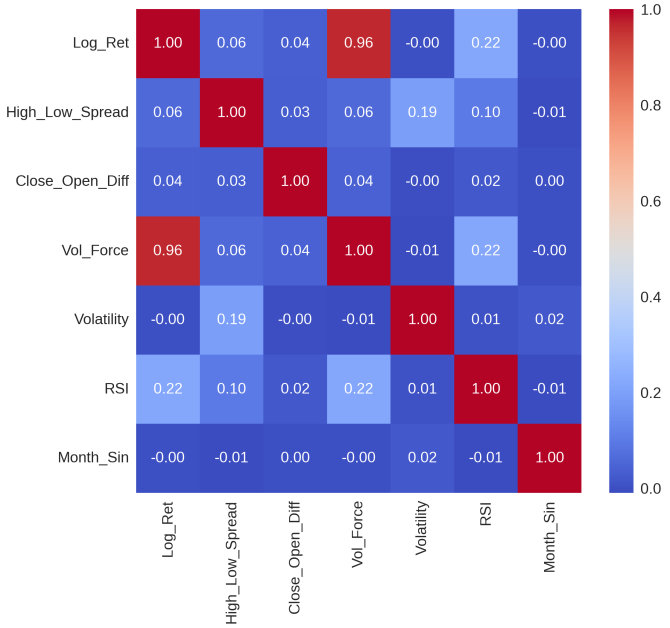


Fig. 4. Correlation matrix showcasing the statistical independence of engineered features.

To address the periodic nature of financial markets, a Cyclical Seasonality Encoding is implemented. The month index is transformed into a continuous sine-wave signal, enabling the model to recognize recurring calendar effects such as fiscal year-end adjustments without the ordinal bias inherent in raw categorical values. This temporal embedding is formally expressed, as in Eq. (7):

$$\text{Month_Sin}_t = \sin\left(\frac{2\pi \times \text{month}}{12}\right) \tag{7}$$

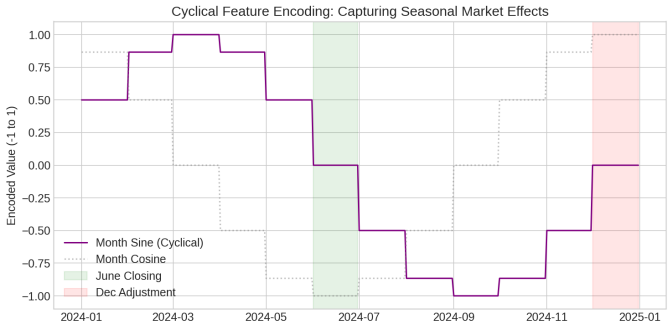


Fig. 5. Cyclical feature encoding for capturing seasonal and periodic market effects.

As illustrated in Fig. 5, this transformation effectively maps seasonal market behaviors, such as June closings and December adjustments, into a normalized periodic space. This multidimensional approach ensures that the input space captures both the geometric morphology of price trends and the underlying temporal regularities of the emerging ecosystem.

D. Proposed Model Architecture: State-AttentNet

To address the limitations of traditional sequential models, this study proposes the State-AttentNet, a hybrid architecture integrating bidirectional temporal learning with an adaptive attention mechanism. The model is specifically designed to decode market regimes by prioritizing high-impact volatility events over stochastic noise.

1) Bidirectional sequential decoding: The core of the architecture is a two-layer Bidirectional LSTM. Unlike standard recurrent units, the Bi-LSTM processes the input feature sequence $X \in \mathbb{R}^{T \times d}$ (where $T = 60$ and $d = 7$) in both forward ($\overrightarrow{h_t}$) and backward ($\overleftarrow{h_t}$) directions. This dual-stream processing allows the network to capture market patterns by analyzing both historical trends and future convergence points simultaneously. The hidden state h_t at any time step t is a concatenation of both directional vectors, as defined in Eq. (8):

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{8}$$

The resulting hidden representation $H = \{h_1, h_2, ..., h_T\}$ encapsulates a robust temporal context of the market dynamics.

2) Adaptive additive temporal attention mechanism: To effectively decode market state,the model employs an additive temporal attention mechanism that dynamically prioritizes critical time steps within the 60-day lookback window.The attention score for each time step is computed, as defined in Eq. (9):

$$e_t = \mathbf{v}^\top \tanh(\mathbf{W}_h h_t + \mathbf{b}) \tag{9}$$

where, $\mathbf{W}_h$ and $\mathbf{b}$ are learnable parameters, and $\mathbf{v}$ is a trainable weight vector.

Instead of treating all historical days equally, this layer assigns a learnable scalar weight α_t to each time step, filtering out stochastic noise while highlighting high-impact volatility

events. The attention scores are normalized using the softmax function, and the final context vector (c) is computed as the weighted sum of the hidden states, as defined in Eq. (10):

$$\alpha_t = \frac{\exp(e_t)}{\sum_\tau \exp(e_\tau)}, \quad c = \sum_{t=1}^{T} \alpha_t h_t \tag{10}$$

The attention mechanism assigns higher weights to time steps that are more informative for regime classification. Practically the model tends to place more emphasis on periods where volatility remains elevated for several days or where the market shows a clear drawdown pattern. Single day jumps that quickly disappear usually get lower weight. This layer is unlike normal self attention where all days are compared with each other daily but this layer assigns each day an importance score and summarizes the entire 60-day window into a single context vector.

3) Classification and output module: The aggregated context vector (c) serves as the input for the final decision-making block. It is propagated through a fully connected Dense Layer (64 units) with ReLU activation to capture non-linear feature interactions. Finally, the output is projected to a probability score using the Sigmoid activation function to classify the market state, as defined in Eq. (11):

$$\hat{y} = \sigma(W_{out} \cdot c + b_{out}) = \frac{1}{1 + e^{-(W_{out} \cdot c + b_{out})}} \tag{11}$$

If the predicted probability $\hat{y} > 0.5$, the system flags a Crash Regime; otherwise, it indicates a stable market condition.

IV. Result Analysis

A. Experimental Setup

The computational experiments were conducted within the Kaggle Notebook cloud environment, leveraging dual NVIDIA Tesla T4 GPUs to accelerate parallel tensor processing and model training. The proposed State-AttentNet architecture was implemented using the PyTorch deep learning framework, while Pandas and Scikit-learn were utilized for data preprocessing and metric evaluation. To ensure robust convergence within the non-stationary DSE environment, the network parameters were optimized using the AdamW algorithm with an initial learning rate of 0.001.

The selection of key hyperparameters was guided by validation performance and stability considerations. The hidden dimension was set to 128 to balance representational capacity and overfitting risk, as lower dimensions resulted in underfitting while higher dimensions did not yield significant improvements. The maximum number of epochs was set to 20 with early stopping based on validation loss, as the model consistently converged within this range. The threshold percentile for dynamic labeling was fixed at 80% based on empirical validation, providing a balanced trade-off between sensitivity to volatility spikes and robustness to noise. Alternative thresholds (e.g., 75% and 85%) were also evaluated but resulted in either increased false positives or reduced crash detection performance. The objective function was minimized using Binary Cross-Entropy. To maintain the structural integrity and temporal fidelity of market signals, the inherent class imbalance was addressed through the Adaptive Temporal Attention mechanism rather than external resampling. This architectural design enables the model to autonomously prioritize low-frequency, high-impact volatility features, effectively learning the dynamics of rare "Crash" regimes directly from original stochastic patterns without introducing synthetic bias. To mitigate optimization stagnation, a ReduceLROnPlateau scheduler was integrated, dynamically decaying the learning rate by a factor of 0.5 if the validation loss failed to improve for 3 consecutive epochs. The dataset was chronologically partitioned into training (2006–2021), validation (2022–2023), and testing (2024–2025) phases, approximating an 80/10/10 distribution. Unlike random shuffling, this temporal split prevents look-ahead bias by ensuring the model utilizes only historical information for training, thereby preserving the predictive integrity essential for financial time-series forecasting. The detailed hyperparameter configuration utilized in this study is summarized in Table II.

TABLE II. Hyperparameter Configuration

Hyperparameter	Value
Optimization Algorithm	AdamW
Scheduler Patience	3 Epochs (Factor=0.5)
Input Sequence (T)	60 Days
Hidden Dimensions	128 (Bi-LSTM)
Batch Size	128
Dropout Rate	0.3
Max Epochs	20

B. Evaluation of Market State Separation

The confusion matrix of State-AttentNet is presented in Fig. 6. The model demonstrates strong separation between stable and stress market regimes, with 66,772 instances correctly classified as Class 0 and 17,871 instances correctly identified as Class 1. Misclassifications remain relatively low, with 3,114 false positives (false alarms) and 3,309 false negatives (missed crash events). The overall accuracy reaches 93%, indicating reliable predictive performance.

From a risk management perspective, false negatives are particularly critical, as they correspond to periods of elevated market risk that remain undetected. The relatively low number of missed crash events suggests that the model effectively captures most volatility-stress regimes. Conversely, false positives represent unnecessary alerts, which may lead to overly conservative decisions or reduced trading activity. The controlled level of false alarms indicates that the model maintains a balanced trade-off between sensitivity and reliability.

Overall, this balance between false positives and false negatives is essential for practical deployment, where both missed risk events and excessive false alarms can have significant financial implications. Furthermore, the high F1-score and the proportional distribution of correct classifications across both classes confirm that the model's decisions are driven by structural volatility signals rather than majority class frequency, validating the efficacy of the internal attention-based weighting.

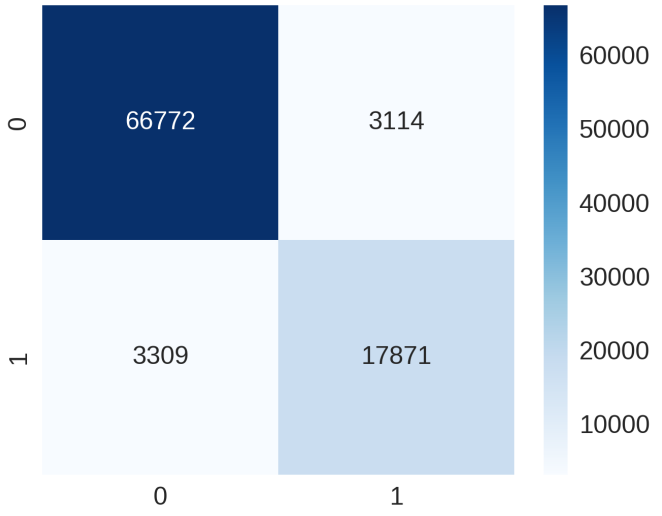


Fig. 6. Confusion matrix of the State-AttentNet model.

C. 30-Day Rolling Accuracy Over Time

Fig. 7 presents the 30-day rolling accuracy of the proposed classifier and provides a transparent view of the behavior of performance over time. The trajectory stays mostly in the upper accuracy region for long stretches indicating that the model maintains reliable state discrimination across varied phases in the market rather than getting a good score using a narrow subset of days. This persistence helps to further the credibility of the reported results because it represents consistent quality of the decision under repeated rolling evaluations. In contrast the plot also shows a limited number of short-lived accuracy drawdowns that cut across the otherwise strong pattern. Such dips are consistent with transition-heavy intervals where volatility clustering and abrupt reversals obscure regime boundaries and reduce separability of features. The rapid return to the high band indicates that the model maintains its strength and clarity once the market dynamics become stable. Operationally, these segments indicate periods when trade signals of confidence have to be handled more conservatively.

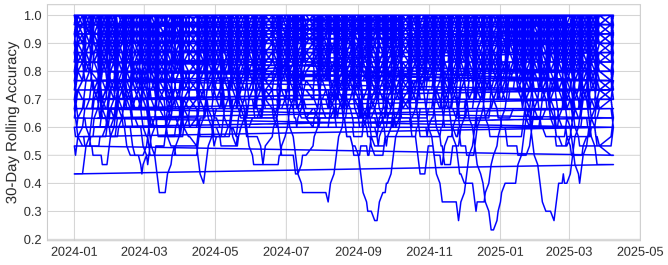


Fig. 7. Rolling accuracy across market phases.

D. ROC–AUC Performance of the State-AttentNet Model

Fig. 8 presents the ROC curve of State-AttentNet with an AUC of 0.97, which suggests high discrimination between the stable and crash regimes over a variety of decision thresholds. The curve indicates that the model is at high true-positives prior to the false-positives rate being relatively low. Interestingly, this finding is derived following a volatility based stable crash labelling framework over which regime markers are set through the rolling volatility shocks and not merely through the direction of prices. This shows how the model can be effective in learning stress-related patterns in the non-stationary and noisy dynamics of the market that can be useful in practical risk-aware regime monitoring in the DSE. A large AUC is significant in the framework of stock markets since regime decisions are threshold sensitive and trading contexts often change because of news, liquidity variations and volatility clusters. The fact that the AUC was always large indicates that the model maintains the quality of rankings of risk scores when a market condition shifts, which is critical to stable alerts and dependable decision support.

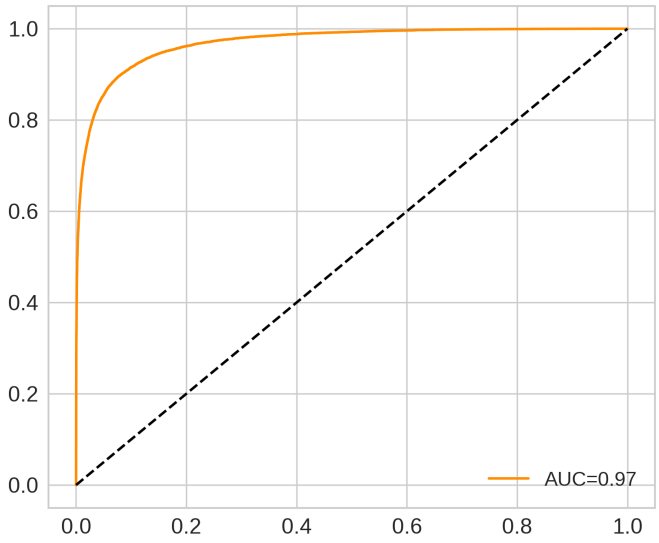


Fig. 8. ROC curve of the State-AttentNet model.

E. Comparative Performance Against Baseline Models with State-AttentNet

Table III demonstrate that State-AttentNet has the best overall performance between the compared baseline models. It attains the highest accuracy at 93.00%, exceeding LSTM with 90.64%, GRU with 91.92%, BiLSTM with 92.51%, TCN with 89.35%, and 1D-CNN with 87.41%. State-AttentNet also has the highest precision of 85.16% and the highest F1-score of 84.77, which means that it has a more balanced trade-off between crashes and false alarms. Despite BiLSTM achieving a marginally higher AUC (0.9709) compared to State-AttentNet (0.9697), the proposed framework demonstrates superior overall performance across key classification metrics. This outcome can be attributed to the temporal attention mechanism, which prioritizes minority-state awareness and enhances contextual interpretability rather than solely optimizing the ranking-based AUC metric. As a result, State-AttentNet achieves a more stable and balanced market-state classification performance, particularly in detecting volatility-driven regime shifts. By contrast, TCN and 1D-CNN have relatively poorer performances, which may indicate that convolution-based architectures are less efficient when it comes to the long-horizon regime of

the DSE. On the whole, these results indicate that in the existing experimental conditions, State-AttentNet can help provide a more accurate and risk-conscious system of market state classification.

TABLE III. PERFORMANCE COMPARISON WITH BASELINE MODELS

Model	Accuracy	Precision	Recall	F1-Score	AUC
LSTM	90.64%	76.19	86.90	81.20	0.9597
GRU	91.92%	79.14	88.64	83.62	0.9683
BiLSTM	92.51%	83.05	85.17	84.10	0.9709
TCN	89.35%	75.21	79.60	77.34	0.9325
1D-CNN	87.41%	71.15	73.25	72.19	0.9055
State-AttentNet	93.00%	85.16	84.38	84.77	0.9697

The outcomes of the ablation clearly show the contribution of each architectural component in Table IV. The elimination of the attention mechanism results in a significant reduction of the performance, especially in recall and AUC, meaning that temporal weighting is very important in determining informative periods of volatility. Likewise, the deletion of bidirectional encoding decreases the model's ability to obtain both forward and backward temporal dependencies, thereby generating less stable predictions and lower overall performance. Removal of both attention and bidirectional structure is the most drastic, and this proves that the two elements work together to increase the representational ability of the model. Conversely, the full model presents the optimal balance among all the measures of evaluation and indicates that when the bidirectional sequence modeling is combined with the temporal attention that is additive, regime persistence and transitioning behavior can be learned more effectively. These results support the design decision to incorporate these elements in the suggested structure.

TABLE IV. ABLATION STUDY OF THE STATE-ATTENTNET COMPONENTS

Model Configuration	Accuracy	Precision	Recall	F1-Score	AUC
w/o Attention	0.9231	0.8559	0.8051	0.8297	0.9619
w/o Bidirectional	0.9141	0.7750	0.8888	0.8280	0.9662
w/o Attn & Bi-directional	0.9165	0.8825	0.7393	0.8046	0.9608
Full Model (State-AttentNet)	0.9300	0.8516	0.8438	0.8477	0.9697

F. Sensitivity Analysis on Lookback Window Size for State-AttentNet

We conducted a sensitivity analysis by changing the input look-back window length T of the model and measuring the impact on predictive performance of Crash to determine the reliability of the proposed State-AttentNet in a high-paced frontier-market condition. Three identical preprocessing pipeline, engineered features, data splits and training procedure were held constant across all of the runs; only the input sequence length T was varied (15, 45, and 60 days). Accuracy, Precision, Recall, F1-score and AUC are reported in Table V.

The results indicate that a lookback window of $T = 60$ provides sufficient temporal context to capture regime persistence and volatility clustering effects. Shorter windows such as $T = 15$ fail to capture extended volatility patterns, resulting in unstable predictions and significantly lower recall for crash detection. Although $T = 45$ improves performance by incorporating more temporal information, it still lacks the full context required to model sustained regime transitions. In contrast, $T = 60$ achieves the best overall performance across all evaluation metrics, suggesting an optimal balance between temporal depth and model stability. These findings demonstrate that the proposed model benefits from longer temporal dependencies and remains robust across moderate variations in the lookback window size.

TABLE V. SENSITIVITY ANALYSIS OF LOOKBACK WINDOW LENGTH

Window Size	Accuracy	Precision	Recall	F1-score	AUC
15	76.87%	80.95	61.46	69.87	0.7753
45	89.66%	75.71	79.13	77.39	0.9399
60	93.00%	85.16	84.38	84.77	0.9697

G. Analysis of Crash Signal with Price Dynamics

Fig. 9 plots the alignment of the State-AttentNet crash alert layer in responding to discontinuous instability in the price path as compared with a response to day-to-day variation. The grey line follows the stock price during 2024-2025 and the red inverted triangles indicate Crash Alert trades that will be made when the risk pattern has been identified by the model as a crash-state signature. The alerts in the early sample are focussed on the sharp decline in late October to mid-November where the price drops rapidly and rebounds with high dispersion. The second thickness cluster is observed at the late November surge and reversal zone and is a sign that extreme upside spikes and subsequent quick pullbacks are experienced by the system as crash like stress since volatility and reversal risk are co-increasing. There are repeated warnings in the last part that is of the sideways yet weak recovery which implies that the classifier is still fragile to unstable consolidation where any small movements can still constitute high downside risk. Generally, the figure proves the point in the study that State-AttentNet makes high-impact regime transitions prior and minimizes the use of stable thresholds by releasing signals primarily during structurally turbulent windows.



Fig. 9. Trade-signal dynamics for crash-risk (volatility-stress) control.

H. Explainability Analysis

To provide a more robust and globally interpretable explanation of the proposed framework, this study employs SHAP instead of relying on a single local surrogate explanation. The SHAP is a more stable and comprehensive measure of model behavior compared to the instance-level interpretation on a local test set, as it quantifies the contribution each instance makes to model behavior on the entire test set.

Volatility and log-return are the two features with the largest impact, as illustrated in Fig. 10 and the maximum mean absolute SHAP values of all the engineered inputs.

This means that the classifier uses mainly the market stress intensity and the dynamics of returns in differentiating between stable and crash regimes. RSI and Vol_Force make secondary contributions, which implies that regime discrimination is also supported by momentum, and trading-activity-related factors. High_Low_Spread, Close_Open_Diff, and Month_Sin, on the other hand, have relatively smaller though not dominating contributions to the prediction process, meaning that they are more complementary than dominant in the prediction process.

These results are in line with the volatility-adaptive labeling design of the research, where the volatility-related signals are supposed to be very informative. The SHAP results however also indicate that the model is not reliant on an individual feature; instead, it combines various engineered market signals in order to enhance the strength of classification. This is significant within a frontier-market environment, where structural instability and noisy price dynamics render regime identification to be sensitive to both dominant and auxiliary market indicators. A single explanation could be over-interpreted; therefore, SHAP values were summed over the entire test set. This offers a dataset-level analysis of feature relevance and minimizes the possibility of making a conclusion based on a single observation. Moreover, the SHAP ranking on the global scale is generally congruent with the analysis of temporal attention, indicating that the model focuses on informative features as well as informative time steps in detecting regime shifts.

In general, the explainability analysis using SHAP justifies the intuitive understanding that the presented framework is responding to the structurally significant market signals instead of reacting to the temporary noise. This enhances transparency and increases the practicality of the model in frontier market risk-conscious regime monitoring.

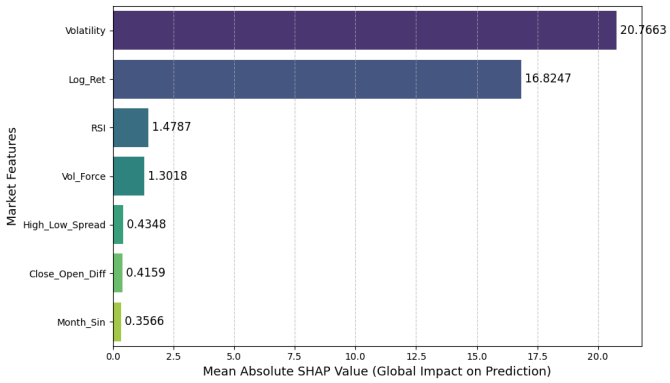


Fig. 10. Global SHAP features importance based on mean absolute SHAP values across the test set.

V. Conclusion

Financial markets are dynamic and constantly moving between stable and stress regimes, requiring the proper identification of the state to have a successful risk management process. Nevertheless, a significant proportion of literature is oriented towards developed markets or is based on predetermined labeling, which may not generalize well to frontier markets with structural volatility, changing liquidity, and time-dependent volatility patterns. In order to fill the gap, this study develops regime detection as an adaptive long-horizon classification problem and implements State-AttentNet to the Dhaka Stock Exchange based on daily data between 1999 and 2025. The framework combines the volatility-based labeling that is dynamic with market-sensitive properties, which enables the definition of stress regimes to change with the market state as opposed to using fixed thresholds.

Empirical data show that State-AttentNet attains an accuracy of 93% and AUC of 0.97, which is better than the LSTM and GRU baselines in the present experimental context. These findings indicate that attention-directed sequence modeling can be effectively used to model regime persistence and transition dynamics in non-stationary and noisy environments. Moreover, the explainability analysis reveals that the model relies on informative lag structures, which is consistent with the meaning of the model capturing meaningful regime shifts not temporary fluctuations. In general, the findings indicate that adaptive labeling combined with attention-based temporal models can offer a viable and understandable regime monitoring framework for the Dhaka Stock Exchange. A limitation of the present study is that the framework was evaluated using data from only the Dhaka Stock Exchange. In addition, class imbalance was addressed solely through the temporal attention mechanism without incorporating conventional strategies such as weighted loss functions or resampling techniques. Consequently, the responsiveness of regime signals to future market behaviour remains unobserved. During real-world implementation, regime alert-based trading decisions can impact liquidity and volatility and can also create looping effects that can change future model inputs. Further development of evaluation within live or paper trading settings will include persistence-based alerting, uncertainty-based decision rules, and controlled alert frequency to maintain stability. It is also possible to apply the framework to real-time monitoring systems with automated data pipelines and rolling retraining to continually update the system as new observations come in. The extensions are further including macroeconomic variables and also microstructure characteristics of the market and also the cross-market generalization with other frontier markets. Increasing the capacity to endure changing market conditions, future studies can incorporate drift detection algorithms to induce retraining and probability calibration methods to perceive model results as risk action measures.

Lastly, one can test the framework in a more applied context by connecting regime indicators to risk-controlling tools, and testing their performance under realistic transaction costs. The general applicability of the proposed approach can be further enhanced by extensions of the multi-class and multi-horizon regime definitions, as well as uncertainty-aware and explainable outputs.

Acknowledgment

The authors would like to express their sincere gratitude to the Southeast University, Dhaka, Bangladesh, for providing the essential laboratory facilities and technical support required to conduct this research.

References

[1] J. D. Hamilton, “A new approach to the economic analysis of nonstationary time series and the business cycle,” *Econometrica*, vol. 57, no. 2, p. 357, Feb. 1989.

[2] A. Ang and G. Bekaert, "International asset allocation with regime shifts," *Review of Financial Studies*, vol. 15, no. 4, pp. 1137–1187, Feb. 2002.

[3] A. Bucci and V. Ciciretti, "Market regime detection via realized covariances," *Econ Model*, vol. 111, p. 105832, Feb. 2022.

[4] Y. Ding, D. Kambouroudis, and D. G. McMillan, "Forecasting realised volatility using regime-switching models," *International Review of Economics & Finance*, vol. 101, p. 104171, Feb. 2025.

[5] A. Poon and D. Zhu, "Do recessions and bear markets occur concurrently across countries? A multinomial logistic approach," *Journal of Financial Econometrics*, vol. 22, no. 5, pp. 1482–1502, Feb. 2024.

[6] A. K. Mishra, J. Renganathan, and A. Gupta, "Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture," *Eng Appl Artif Intell*, vol. 133, p. 108223, Feb. 2024.

[7] S. Kumar, A. Rao, and M. Dhochak, "Hybrid ML models for volatility prediction in financial risk management," *International Review of Economics & Finance*, vol. 98, p. 103915, Feb. 2025.

[8] X. Song, W. Chen, and J. Lu, "Research on financial stock market prediction based on the hidden quantum markov model," *Mathematics*, vol. 13, no. 15, p. 2505, Aug. 2025.

[9] Z. Song, H. S. H. Tsang, R. T. C. Hsung, Y. Zhu, and W. L. Lo, "From market volatility to predictive insight: An adaptive transformer–RL framework for sentiment-driven financial time-series forecasting," *Forecasting*, vol. 7, no. 4, p. 55, Feb. 2025.

[10] Md. A. K. Azad *et al.*, "Stock price forecasting with temporal-feature attention and graph SAGE on multimodal financial data," in *2025 International Conference on Intelligent and Secure Engineering Solutions (CISES)*, IEEE, 2025, pp. 718–723.

[11] M. H. Begum *et al.*, "Diverse minds think alike: A SHAP-guided RNN for stock price forecasting with smart SMA optimization," in *2025 3rd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, IEEE, 2025, pp. 1–5.

[12] A. K. M. Masum *et al.*, "A topology-driven CNN framework for stock price prediction using news sentiment and market dynamics," in *2025 International Conference on Intelligent and Secure Engineering Solutions (CISES)*, IEEE, 2025, pp. 712–717.

[13] A. Koukorinis, G. W. Peters, and G. Germano, "Generative-discriminative machine learning models for high-frequency financial regime classification," *Methodol Comput Appl Probab*, vol. 27, no. 2, p. 36, Feb. 2025.

[14] P. Nystrup, H. Madsen, and E. Lindström, "Dynamic portfolio optimization across hidden market regimes," *Quant Finance*, vol. 18, no. 1, pp. 83–95, Feb. 2018.

[15] M. D. A. S. Sunny, "Dhaka Stock Exchange Historical Data (1999-2025)," *Harvard Dataverse*, 2025. [Online]. Available: https://doi.org/10.7910/DVN/XIFYT1

[16] J. D. Hamilton and R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *J Econom*, vol. 64, no. 1–2, pp. 307–333, Feb. 1994.

[17] S. F. Gray, "Modeling the conditional distribution of interest rates as a regime-switching process," *J financ econ*, vol. 42, no. 1, pp. 27–62, Feb. 1996.

[18] S. Blazsek, D. Kong, and S. R. Shadoff, "Within-regime volatility dynamics for observable- and Markov-switching score-driven models," *Financ Res Lett*, vol. 73, p. 106631, Mar. 2025.

[19] P. Zegadło, "Identifying bull and bear market regimes with a robust rule-based method," *Res Int Bus Finance*, vol. 60, p. 101603, Feb. 2022.

[20] S. Wang, R. Gupta, and Y. J. Zhang, "Bear, bull, sidewalk, and crash: The evolution of the US stock market using over a century of daily data," *Financ Res Lett*, vol. 43, p. 101998, Feb. 2021.

[21] J. M. Maheu, T. H. McCurdy, and Y. Song, "Bull and bear markets during the COVID-19 pandemic," *Financ Res Lett*, vol. 42, p. 102091, Feb. 2021.

[22] W. N. Lai, "Detecting stock market regimes from option prices," *Operations Research Letters*, vol. 50, no. 3, pp. 260–267, Feb. 2022.

[23] C. Erlwein-Sayer, S. Grimm, A. Pieper, and R. Alsaç, "Forecasting corporate credit spreads: Regime-switching in LSTM," *Econom Stat*, Feb. 2023.

[24] F. Haase and M. Neuenkirch, "Predictability of bull and bear markets: A new look at forecasting stock market regimes (and returns) in the US," *Int J Forecast*, vol. 39, no. 2, pp. 587–605, Feb. 2023.

[25] P. Chen, D. Yi, and C. Zhao, "Trading strategy for market situation estimation based on hidden Markov model," *Mathematics*, vol. 8, no. 7, p. 1126, Feb. 2020.

[26] I. F. C. Battazza, C. M. de Oliveira Rodrigues, and J. F. L. de Oliveira, "A framework for market state prediction with ontological asset selection: A multimodal approach," *Applied Sciences*, vol. 15, no. 3, p. 1034, Feb. 2025.

[27] T. Mudarisov, R. V. State, Z. Kraussl, A. Yakubov, and T. Petrova, "Cross-sector market regime forecasting with LLM-augmented news analysis," in *Proceedings of the 5th ACM International Conference on AI in Finance*, ACM, 2024, pp. 461–468.

[28] M. X. Hanauer and T. Kalsbach, "Machine learning and the cross-section of emerging market stock returns," *Emerging Markets Review*, vol. 55, p. 101022, Jun. 2023.

[29] J. Grudniewicz and R. Ślepaczuk, "Application of machine learning in algorithmic investment strategies on global stock markets," *Res Int Bus Finance*, vol. 66, p. 102052, Oct. 2023.

[30] J. H. Liou, Y. T. Liu, and L. C. Cheng, "Price spread prediction in high-frequency pairs trading using deep learning architectures," *International Review of Financial Analysis*, vol. 96, p. 103793, Nov. 2024.

[31] V. Pakštaitė, E. Filatovas, M. Juodis, and R. Paulavičius, "Bitcoin price regime shifts: A Bayesian MCMC and hidden Markov model analysis of macroeconomic influence," *Mathematics*, vol. 13, no. 10, p. 1577, Feb. 2025.

[32] L. Wang, S. An, Z. Dong, X. Dong, and J. Li, "Early warning of regime switching in a financial time series: A heteroskedastic network model," *PLoS One*, vol. 20, no. 10, p. e0333734, Feb. 2025.

[33] Y. Shu, C. Yu, and J. M. Mulvey, "Dynamic asset allocation with asset-specific regime forecasts," *Ann Oper Res*, vol. 346, no. 1, pp. 285–318, Feb. 2025.

[34] R. Gupta, S. Kapoor, H. Gupta, and S. Natesan, "A forest of opinions: A multi-model ensemble-HMM voting framework for market regime shift detection and trading," *Data Science in Finance and Economics*, vol. 5, no. 4, pp. 466–501, 2025.

[35] M. Davidovic and J. McCleary, "News sentiment and stock market dynamics: A machine learning investigation," *Journal of Risk and Financial Management*, vol. 18, no. 8, p. 412, Feb. 2025.

Cognitively Aligned Assessment Item Generation with Open-Source LLMs: A Comprehensive Evaluation on LearningQ

Mahmoud Badry[1], Walaa Medhat[2], Shereen A. Taie[3], Asmaa Hashem Sweidan[4]
Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum, Egypt[1,3,4]
Faculty of Computers and Artificial Intelligence, Banha University, Qalyubiyya, Egypt[2]

***Abstract*—Automated generation of high-quality educational assessment items is still difficult, especially when it comes to higher-order cognitive skills. Although Large Language Models (LLMs) show promise, their structural validity and cognitive alignment are limited. This study systematically evaluates fine-tuning open-source LLMs using an enriched LearningQ dataset that includes Bloom's cognitive labels and evidence. The results show a clear performance contrast. Qwen2.5-3B-Instruct displays the best semantic reasoning, while Llama-3.2-3B shows better structural adherence, achieving a 94.9% validity rate and full compliance against answer leakage while maintaining high question validity. In contrast, older encoder-decoder models like FLAN-T5-XL do not generate valid questions. The study finds that small- to medium-sized instruction-tuned models, backed by strong data engineering, are successful at developing scalable, cognitively well-aligned assessment items.**

***Keywords*—Automated question generation; Large Language Models; educational assessment; Bloom's Taxonomy; Parameter-Efficient Fine-Tuning; LearningQ**

I. Introduction

The application of Large Language Models (LLMs) in educational technology has transformed automated content generation. This is particularly true for creating high-quality assessment items that assess not just factual recall but also higher-order thinking (cognitive) skills such as analysis, application, and synthesis. Traditional automated question generation (AQG) methods were based on rule-based systems or early neural models, which had limitations in scope and scalability. Instruction-tuned large language models can generate a wide range of effective questions at scale. However, models that are generally pre-trained often produce results that lack educational depth or fail to follow the strict structural rules required for standardized assessments.

Cognitive alignment refers to how well an assessment item measures the cognitive processes intended by a given learning outcome within the field of educational measurement theory. The idea of cognitive alignment is very similar to that of construct validity and alignment validity within the realm of test creation. Thus, assessing the cognitive alignment of test creation involves more than simply looking at linguistic similarities. In this study, an "assessment item" refers to a "single educational question" designed to evaluate a specific learning objective or cognitive level.

A major limitation in current AQG research is the reliance on datasets meant for machine reading comprehension, like SQuAD [1] or RACE [2]. Although these datasets are large, they mainly focus on extractive fact retrieval ("Remembering"), which does not capture the details of educational assessment. Assessment items need to evaluate specific cognitive levels, as defined by frameworks such as Bloom's Taxonomy. Additionally, valid assessment items must meet strict requirements. For example, the questions should be interrogative, the answer should not be included in the question, and premature explanations should not be given. General-purpose LLMs often violate these rules in zero-shot settings.

Beyond structural validity, high-quality educational items must also show a variety of assessment formats and Intended Learning Outcomes (ILOs). This includes generating items across multiple item types—such as conceptual, analytical, application-based, and synthesis-oriented prompts—and aligning each item with a specific learning objective rather than producing generic fact-recall items. Without enforcing this alignment, models often revert to simple cognitive tasks, which limits their usefulness in real teaching and evaluation situations.

To address these challenges, this study examines the ability of open-source LLMs to create standardized assessment items using LearningQ [3], a large dataset derived from real educational platforms, including TED-Ed and Khan Academy. Unlike datasets designed to find facts, LearningQ has many questions that require higher-order thinking. This quality makes it a suitable testing ground for developing questions that align with cognitive principles. This study focuses on assessing various open-source LLM models using a unified, cognitively aligned pipeline, rather than comparing them with existing state-of-the-art AQG systems. This is to establish the influence of models on cognitive alignment.

This study presents a detailed fine-tuning and evaluation framework to close the gap between raw educational content and usable assessment items. The proposed approach employs a comprehensive data enrichment pipeline that initially improves raw LearningQ data by adding synthetic cognitive labels and emphasized passage evidence, utilizing Qwen2.5-7B-Instruct [4]. Next, several models are adjusted using Parameter-Efficient Fine-Tuning (PEFT) through Low-Rank Adaptation (LoRA) [5]. This includes both contemporary decoder-only architectures, such as Llama-3.2-3B [6] and Qwen2.5-3B, and traditional encoder-decoder models like FLAN-T5-XL [7].

This research makes several significant contributions:

1) A Cognitive-Aligned Data Pipeline: that adds passage highlighting and Bloom's Taxonomy labels to educational datasets. This enables models to discover the relationship between cognitive targets and segments of the source text.
2) Comparative Model Analysis: which evaluates six different LLM architectures with 3 billion to 7 billion parameters. It looks at their performance not only on lexical overlap using BLEU and ROUGE but also on semantic preservation through BERTScore and pedagogical intent.
3) Insights for AQG: show that modern instruction-tuned LLMs, when combined with careful data engineering and fine-tuning, can consistently produce assessment items that align with cognitive goals, outperforming traditional architectures.

Our results indicate a distinct performance gap between contemporary instruction-tuned architectures and earlier models. This demonstrates that with proper data engineering and fine-tuning, small-to-medium LLMs can consistently function as tools for generating automated, cognitively aligned assessments. This may assist educators in developing standardized assessments.

The remainder of this study is organized as follows: Section II reviews related state-of-the-art literature and identifies the main challenges of current research. Section III describes the LearningQ dataset. Section IV presents the phases of the proposed approach. Section V and Section VI discuss the evaluation criteria and results based on the enriched LearningQ dataset. Section VII discusses the implications of the findings. Section VIII presents conclusions and discusses future work.

II. Related Work

This section presents related work on relevant datasets, methods for generating questions using LLMs, and applications of the LearningQ dataset [3].

A. Datasets for Question Generation

Significantly, the establishment of robust datasets has been critical in the development of machine reading comprehension and question generation. Early datasets such as SQuAD [1], which have over 100,000 crowd-sourced questions associated with corresponding passages on Wikipedia, with answers being restricted to a particular text span, have been instrumental. Although such datasets have been effective for extractive question answering, they have been primarily focused on factual recall. To address this issue, RACE [2], a dataset of multiple-choice questions created by experts and based on English examination papers for middle and high school students has been proposed. Unlike SQuAD, the answers in RACE cannot be extracted; they require a thorough level of reasoning and understanding of a passage.

The attempt to bridge QA studies and educational assessment has been realized in the EduQG dataset, released in 2023 [8], which contains a collection of 3,397 questions in multiple formats, all gathered from OpenStax textbooks. EduQG also introduced a new type of sentence-level annotation and included a subset of 903 questions annotated with their corresponding level of cognitive complexity according to Bloom's Taxonomy.

Following this trend, EduQuest, released in 2024 [9], has widened the field of application of question answering in education by collecting over 76,000 instructional documents and 68,000 questions from sources such as MIT OpenCourseWare, OpenStax, and Khan Academy. An important aspect of EduQuest is the addition of metadata labels, such as the corresponding cognitive complexity level and pedagogical labels, which are essential for applications that involve higher-level question generation beyond simple recognition and recall.

B. Educational Assessments and LLM-Based Question Generation

Recent advances in Large Language Models (LLMs) have motivated the use of these models to facilitate the automation of educational assessment design. Lamsiyah et al.'s (2024) [10] study presents an approach to automatic question generation using the RLLM-EduQG model, which fine-tunes the FLAN-T5 model on a hybrid objective function combining cross-entropy and reinforcement-learning rewards based on BLEU and semantic similarity scores. This approach addresses exposure bias and improves both syntactic and semantic generation performance. Another study on automatic question generation was proposed by Hasan et al. (2024) [11]. It presented an Automatic Question & Answer Generation (AQAG) framework based on the Llama-2 model and uses unsupervised prompts on the RACE corpus to generate multiple-choice and fill-in-the-blank question types.

Regarding cognitive alignment, Zhuge et al. (2025) [12] introduced TwinStar. It is a dual-LLM system in which one model generates questions and another. It was fine-tuned on ChatGPT-6B to assess alignment at cognitive levels. This refining process better preserved cognitive targets and ensured content relevance compared to models such as GPT-4 and Bard.

C. Applications of the LearningQ Dataset

LearningQ has gained popularity in recent research studies due to its scope, instructional value, and diverse questions. Focal (2023) [13], an end-to-end automated construction of an evaluation pipeline, utilized LearningQ to fine-tune its evaluator model. The study showed promising outcomes in enhancing the logical coherence and instructional value of questions generated by LLMs. Moreover, García-Méndez [14] (2025) emphasized LearningQ's value in training models for generating questions as well as automated evaluations, citing previous studies by Moore et al. (2022) and Bhat et al. (2022) on training models to evaluate the quality of questions generated by students with the aid of LearningQ.

In a study of multimodal learning, Stamatakis et al. (2025) [15] analyzed LearningQ, a subset of TED-Ed and Khan Academy, to assess Vision-Language Models (VLMs) such as Video-LLaMA and PG-Video-LLaVA. The study showed that when generating contextually relevant educational questions from video inputs, VLMs trained or fine-tuned on LearningQ achieved a significant improvement over zero-shot VLMs that generated irrelevant queries.

D. Cognitive Alignment and LLM Evaluation

Some studies also used cognitive models, such as Bloom's taxonomy, to evaluate LLM performance. Several LLMs, such as Llama-2, Mistral, and GPT-4, were assessed by Scaria et al. (2024) [16] using cognitive prompt techniques, such as the Chain of Thought. According to this study, LLM models typically perform well on lower-order thinking tasks such as "Remember" or "Understand", but struggle to produce reliable results on higher-order thinking tasks such as "Create". Another study by Huber and Niklaus (2025) [17] assessed well-known benchmark tests, such as MMLU and BIG-Bench Hard, and classified them according to Bloom's taxonomy levels. According to this study, there were very few instances of higher-order thinking tasks like "Evaluate" or "Create" in these benchmarks, while it found a high frequency of lower-order thinking tasks.

Most of the literature reviewed has focused on creating automated questions with the assessment AQG. These studies offer insights related to cognitive alignment. However, only a few studies investigate cognitive alignment directly. These have issues, including poor structural validity, poor cognitive alignment, and poor adherence to assessment constraints in previous AQG approaches. This study addresses these problems using the proposed approach.

III. Dataset: LearningQ

A. Dataset Description

LearningQ [3] is a large dataset created specifically to generate and analyze educational questions, with a special emphasis on learning- and assessment-oriented questions. This is unlike other datasets, such as SQuAD [1] and RACE [2], which have mostly focused on information retrieval questions. LearningQ comprises more than 230,000 document-question pairs across domains such as science, humanities, and mathematics.

LearningQ data was collected from two major online learning platforms: **1) TED-Ed** (https://ed.ted.com/) consists of questions created by instructors linked to educational video lectures, characterized by high quality and a well-structured educational framework. **2) Khan Academy** (https://www.khanacademy.org/) set includes questions prepared by learners and posted in the comments section of videos and articles. To reduce noise, the dataset creators used a convolutional neural network (CNN)-based classifier to filter out irrelevant content while retaining learning-relevant questions, achieving 80.5% accuracy.

LearningQ consists of videos and article documents coupled with educational questions. It features documents that are much longer than standard benchmarks and covers the revised Bloom's taxonomy levels with distributions of recall (31%), comprehension (36%), application (9%), analysis (11%), evaluation (2%), and creation (1%), along with nearly 10% of unknown/irrelevant content generated by the learner, as shown in Table I. For clarity, Table I presents the datasets by content type (video lectures and articles) obtained from the platforms TED-Ed and Khan Academy. To better understand the structural properties of LearningQ, the proposed framework conducted an in-depth statistical and schema-level inspection of the JSON files using Python utilities. The TED-Ed subset has a standardized, instructor-designed structure, unlike Khan Academy, where a high level of variability was observed, with a range of 1 to 42 questions per document, a mean of 12.44, and a standard deviation of approximately 8.71. LearningQ covers a modified version of Bloom's Taxonomy, covering a wide range of cognitive abilities from basic recall to complex thinking, offering a high level of cognitive and structural variation (see Table II).

TABLE I. Statistics of the LearningQ Dataset

Data Source	Content Type	# Docs	# Qs	Avg Q/Doc
TED-Ed	Video Lectures	1,102	7,612	6.91
Khan Academy	Video Lectures	7,924	201,273	25.40
Khan Academy	Articles	1,815	22,585	12.44
Total	**Mixed**	**10,841**	**231,470**	**21.3**

TABLE II. Question Examples of Different Bloom's Revised Taxonomy Levels in LearningQ Dataset.

Bloom Level	Example
Remembering	What is a negative and positive feedback in homeostasis?
Understanding	Why do some plankton migrate vertically?
Applying	If I double the area and take half of the fraction, do I get the same result?
Analyzing	Why are cities like London, Tokyo, and New York facing shortages in burial ground space?
Evaluating	Will all the cultures merge into one big culture, due to the fading genetic distinctions?
Creating	Can somebody please explain to me what marginal benefits are and give me some examples?

While LearningQ is presented as covering a range of cognitive levels, the downloaded dataset does not include Bloom's Taxonomy labels for individual questions. Thus, the provided distribution represents only a descriptive understanding of the data. Therefore, cognitive-level labeling has been newly applied in this study using a language model–based classifier.

IV. Methodology

This section describes the proposed framework, which applies a structured process to transform raw educational data into high-quality assessment items. As shown in Fig. 1, the framework includes four clear phases: 1) Data Preprocessing and Enrichment, which involves raw LearningQ data [3]; 2) Prompt Engineering, which organizes the data into instruction-tuned structures; 3) Parameter-Efficient Fine-Tuning (PEFT), which uses Low-Rank Adaptation (LoRA) to fine-tune LLMs for the educational domain; and 4) Comprehensive Evaluation, which includes both semantic measures and evaluation constraints.

A. Data Preprocessing and Enrichment

It plays an important role, significantly impacting the overall performance of subsequent processes. Consequently, the input data pass through a series of steps to clean, normalize, and semantically validate them. The preprocessing stage involves the following steps:

1) Dataset Cleaning: This step includes nested JSON object analysis, text normalization, handling of invalid

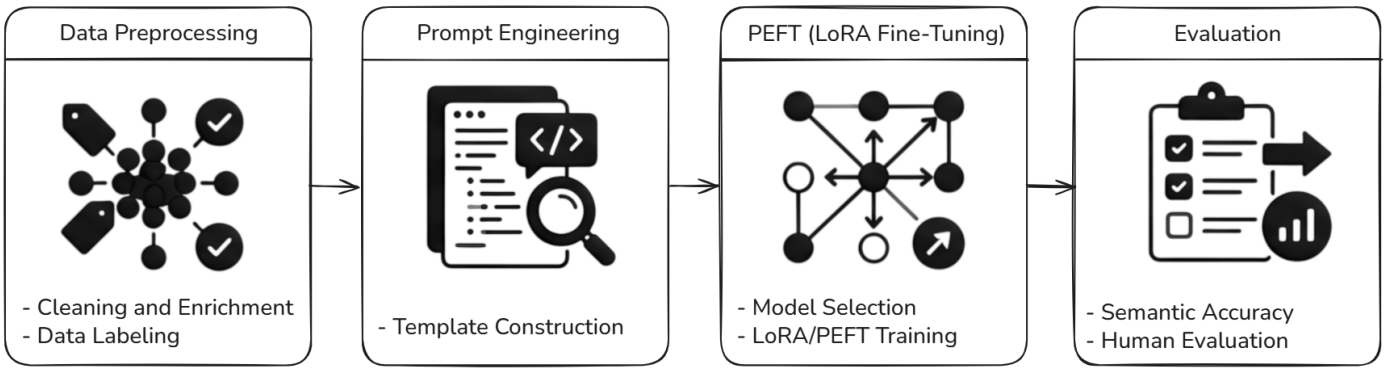


Fig. 1. Methodology pipeline

rows, and resolving inconsistencies between multiple sources. The output of this is a flat table, Table III. Specifically, any invalid entries (e.g., empty passages, questions, or malformed data) were filtered out. The text data was normalized by converting to lowercase and trimming, and any samples with invalid question-passage alignment were filtered.

2) Restructuring into One-Question-Per-Row: To ensure that the supervision during the training is consistent, the data is transformed into the following structure: "| ID | Title | Subject | Text | Question |".
3) Cognitive Level Classification: Since LearningQ does not provide cognitive labels, each question in LearningQ was labelled by following the taxonomy of Bloom. For this experiment, Qwen2.5-7B-Instruct [4] was used because it was observed to perform well at following instructions, was efficient in 8-bit deployment, and exhibited stability in differentiating between nearby cognitive levels. A new column, "Cognitive Level", was added to the schema.
4) Highlighting Relevant Passage Segments: A second LLM processing step (Qwen2.5-7B-Instruct) is utilized to pinpoint the relevant passage segments needed to answer the questions. These segments are marked with the HTML tag '<h>...</h>. The purpose of highlighting was to improve alignment between the passage and the question. This provides explicit contextual grounding and facilitates easier interpretation during evaluation.
5) Verify reliability: To verify the reliability of the supervision signals produced by the LLM for training purposes, a random set of 150 supervision signals was manually checked for their validity. The agreement between the two human evaluators for the Bloom-level labels was substantial, with a Cohen's κ of 0.79. Furthermore, the highlighted evidence segments were found to be 87% accurate according to human judgment.
6) Balancing Cognitive-Level Distribution: After labelling and highlighting, the distribution of classes at different Bloom levels was measured. To ensure a balanced distribution, classes that were overrepresented were downscaled, while underrepresented classes were upscaled.

TABLE III. DATASET CLEANING: INITIAL SHEET OUTPUT

Field	Description
ID	Source document identifier
Title	Document title
Subject	High-level semantic category
Text	Passage/content body
Questions	List of associated questions

Template Components

@@TASK@@
Generate exactly ONE assessment question.

@@SUBJECT@@
<Subject>

@@COGNITIVE_LEVEL@@
<Cognitive Level>

@@PASSAGE@@
<Text>

@@CONSTRAINTS@@
The question must be answerable using ONLY the passage
Do NOT include the answer
Do NOT include explanations
Output the question only

Fig. 2. Prompt template components.

B. Prompt Template Construction

To ensure consistent supervision during fine-tuning and minimize stylistic variability across training instances, each enriched record is converted into a unified instruction–response template Fig. 2, including the topic, knowledge level, and highlighted passage, with strict output constraints. Task specifications ("Create only one assessment question") guide the creation of a single item and establish a consistent pattern for fine-tuning. Metadata and <h>-tagged evidence provide context, while the output constraints — preventing explanations, avoiding answer leakage, and requiring only one question — ensure that assessment items are standardized, organized, and cognitively aligned.

C. Model Loading and LoRa Fine-Tuning

Pairs of prompt and response examples were used to fine-tune six open-source large language models using Parameter-Efficient Fine-Tuning (PEFT) [5]. To address resource constraints in educational natural language processing, the Low-Rank Adaptation (LoRA) technique was applied. This approach introduces a small number of trainable parameters into the main transformer layers while keeping the base model weights fixed, enabling efficient domain adaptation without full end-to-end training. The entire process is illustrated in Fig. 3.

1) Model Selection and Architectural Diversity: To ensure a comprehensive evaluation across architectures, six models were fine-tuned under identical experimental conditions: Llama-3.2-3B-Instruct [6], Qwen2.5-3B-Instruct [4], Mistral-7B-Instruct [18], MediPhi-Instruct [19], Gemma-3-4B [20], and FLAN-T5-XL [7]. These models span a spectrum of decoder-only, encoder–decoder, and multimodal architectures with 3B–7B parameters, allowing us to evaluate differences in model behavior under a uniform task formulation.

2) Quantized Model Loading: All models were loaded in 8-bit precision, consistent with the training script. This reduced VRAM usage by approximately 60%, enabling fine-tuning on modest GPU hardware such as Google Colab and Kaggle T4 instances. Despite the compression, 8-bit quantization preserves the representational depth required for reasoning over instructional prompts and highlighted passages.

3) Dataset Formulation and Preprocessing: The dataset was divided into subsets for training (80%), validation (10%), and testing (10%), subsets and organized into unified objects to facilitate batch processing. Input sequences were tokenized, using a structural separator (@@PASSAGE@@) to separate instructions from paragraph content, and the marker "### OUTPUT": to indicate the target response. Sequences were truncated to fit the model's context window, and selective masking of tokens (-100) was applied to the instructions, paragraph content, and padding to ensure that the loss function was optimized only for the generated question outputs.

4) Fine-Tuning Configuration and Execution: To optimize the pre-trained models for the target task while minimizing computational overhead, Low-Rank Adaptation (LoRA) was employed. The adapter configuration targeted all key linear layers within the transformer architecture, including the query, key, value, and output projections (q_proj, k_proj, v_proj, o_proj), as well as the feed-forward gate, up, and down projections. The hyperparameters were set to r=8, alpha=16, and a dropout rate of 0.05. This configuration introduces fewer than 0.3% additional trainable parameters, enabling efficient adaptation for task-specific behaviors—such as cognitive-level alignment and strict output formatting—without the resource intensity of full-model fine-tuning. Model training was performed with 8-bit quantization and mixed precision (FP16), a batch size of 1 per device, gradient accumulation every 4 steps, and a learning rate of 2e-4 for 3 epochs. The input sequences were limited to 2048 tokens, with a structured prompt format and masking, so that the loss was calculated only on the output tokens. The preprocessing included data normalization and filtering out invalid entries. Every 100 training steps, validation was used to evaluate the model's performance. During this period, the effectiveness of the network was evaluated through perplexity, defined as $e^{\text{eval_loss}}$.

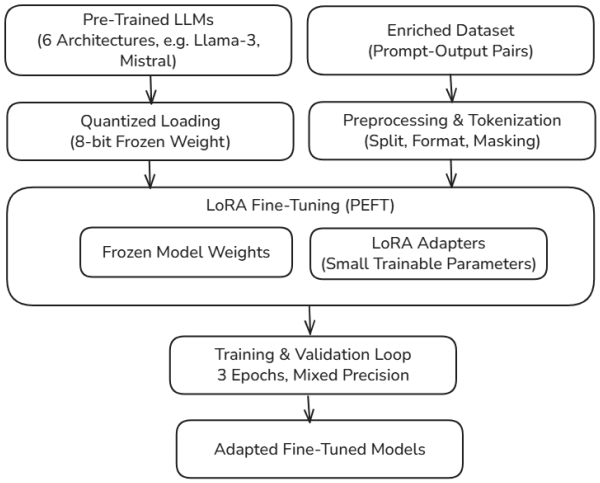


Fig. 3. The end-to-end LoRA fine-tuning framework across diverse model architectures.

D. Inference and Comprehensive Evaluation

Following fine-tuning, a series of evaluations was performed using a controlled inference process, that simulated the exact conditions under which the model was trained. In this process, the entire prompt template was recreated, including the task directive, subject, cognitive level, highlighted passage, and delimiter, to ensure no distributional shift occurred in the model's outputs. The outputs generated by a deterministic decoding (greedy, non-sampling) were then normalized to assess their structural validity, and ensure that only questions were evaluated.

To ensure reproducibility, experiments were performed with a fixed random number generator seed (42). The input sequences were limited to 2048 tokens, with the prompt and passage text truncated as needed. During inference, deterministic decoding was used (greedy decoding with a maximum of 64 tokens (max_new_tokens = 64)). The output was normalized, and the constraints were verified.

V. Evaluation Metrics

A thorough evaluation of the quality of generated questions in the education domain was conducted using multidimensional evaluation metrics. Unlike traditional text-generation applications, cognitive alignment is not an observable phenomenon that can be measured by a single automatic metric. Rather, it requires a multidimensional measurement approach, in which each measure reflects an element of the construct in question. Table IV shows how the components of cognitive alignment relate to the proxies used to measure them.

It is important to understand that the commonly used action verbs connected with Bloom's taxonomy ("define", "explain", "analyze") are very poor indicators of the cognitive process involved. Different cognitive processes can be represented by the same action verb based on their levels of difficulty. For example, "Explain what photosynthesis is" means the use of knowledge, while "Explain how the environmental changes

TABLE IV. MAPPING OF COGNITIVE ALIGNMENT COMPONENTS TO EVALUATION METRICS

Construct Component	Proxy Metric
Semantic Similarity	BERTScore, SBERT
Lexical and Structural	BLEU, ROUGE, METEOR
Structural Validity	is_question, length_ok
Assessment Integrity	no_answer_word, no_expl_word
Cognitive Intent Alignment	Human Evaluation

affect the rate of photosynthesis" requires analysis. Therefore, it is impossible to rely solely on keyword categorization to determine the cognitive process level.

A. Semantic Similarity Metrics

These metrics evaluate how well questions generated from a reference question preserve the original meaning and intent, despite possible variations in wording.

1) BERTScore (F1) [21]: It is used to evaluate semantic similarity based on the comparison of contextual word representations of generated candidates and reference texts. In contrast to n-gram-based metrics, BERTScore is robust to paraphrasing. Higher scores indicate greater semantic similarity between generated questions and the ground truth.

2) Sentence-BERT (SBERT) Cosine Similarity [22]: In the Sentence-BERT model, the cosine similarity is used to measure the similarity between the sentence embeddings of the generated output and the reference text.

B. Lexical and Structural Metrics

It measures Conventional n-gram-based metrics evaluate the generated text's language quality and structure in relation to the reference questions.

1) BLEU (Bilingual Evaluation Understudy) [23]: It assesses how similar the output and reference are in terms of n-gram overlap. In generative tasks, the BLEU metric is usually lower; however, a higher BLEU score denotes higher output precision.

2) ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [24]: The F1 score of the ROUGE-L system has been used for the assessment of the longest common subsequence that exists between the generated candidate and the reference texts.

3) METEOR [25]: It uses stem, synonym, and word order similarities to measure similarities. This metric is generally more flexible with respect to linguistic variation than BLEU. It's also known to correlate strongly with human judgments of word similarity.

C. Task-Specific Constraint Validation

In addition to linguistic similarities, it is important that the generated output is functionally correct with respect to the assessment. Therefore, a constraint-checking module is implemented using heuristic algorithms.

1) Question Validity (is_question): The first heuristic evaluates if the output complies with the syntax of a question. It does this through the presence of a question mark and interrogative sentence structure.

2) Answer Leakage (no_answer_word): Another important failure mode in question generation involves including the answer within the answer (for example, "The answer is."). It ensures that there is no answer leakage within the generated question.

3) Explanation Prohibition (no_expl_word): The models are expected to produce questions only. There is no need for any explanation. The purpose of the check is to ensure the production of 'pure' questions without any explanation.

4) Length Constraints (length_ok): This metric checks whether the generated question falls within an appropriate range of characters or tokens, preventing the generation of either overly short or overly long questions.

D. Cognitive Intent Alignment

For the task of determining the relevance of the generated questions based on the specified Bloom's Taxonomy level, a human evaluation approach was employed. A random sample of questions generated was evaluated independently by two human annotators who had the passage, the required cognitive level, and the generated question available for consideration and were asked to check if the generated question matched the desired cognitive level. The basis of the evaluation is educational measurement theory, according to which an evaluation by human experts is the most reliable method for measuring cognitive processes and their alignment to learning goals. This human evaluation was conducted intentionally because automatic measurements alone do not capture all aspects of pedagogical value or cognitive alignment across Bloom's levels.

VI. RESULTS

A. Overall Performance Evaluation

A comprehensive set of semantic, lexical, and constraint-based metrics was used to assess six refined large language models (LLMs). Llama-3.2-3B, Qwen2.5-3B, Mistral-7B, MediPhi, Gemma-3-4B, and FLAN-T5-XL were the models evaluated. Their ability to create assessment items that meet cognitive expectations was evaluated. A summary of the performance comparison across all important metrics is presented in Table V. Table VI shows qualitative differences in model-generated assessment items.

B. Semantic Similarity Analysis

Embedding-based metrics are used to evaluate how well the generated questions maintain the ground truth's educational intent.

- Qwen2.5-3B achieved the highest BERTScore F1 (0.8275), indicating it was most effective at capturing the underlying semantic meaning of the reference questions, even when the specific phrasing differed.
- Mistral-7B demonstrated superior performance in Sentence-BERT (SBERT) cosine similarity (0.3272), suggesting a strong alignment in overall sentence-level intent.

- Llama-3.2-3B and MediPhi remained competitive with BERTScores above 0.80, whereas Gemma-3-4B significantly underperformed (0.447), indicating a failure to generate semantically relevant content.

C. Lexical and Structural Fidelity

In terms of exact lexical matching and structural coherence:

- Llama-3.2-3B outperformed all other models in BLEU (0.1131) and METEOR (0.0989) scores. This suggests that Llama-3.2-3B most frequently generated n-grams and stems that matched the ground truth references.
- Mistral-7B achieved the highest ROUGE-L F1 (0.0914), marginally outperforming Qwen2.5-3B (0.0904), indicating strong capabilities in preserving the longest common subsequences and structural phrasing of the assessment items.
- FLAN-T5-XL failed in these categories, scoring 0.0 across BLEU, ROUGE-L, and METEOR. Qualitative inspection revealed that this model often generated short, non-question phrases (e.g., "The need for rights in government") rather than valid interrogatives.

D. Task-Specific Constraint Compliance

The usefulness of the models for educational purposes was evaluated based on the models' compliance as follows:

- Question Validity (is_question): The highest validity rate of 94.9% was recorded by Llama-3.2-3B. MediPhi followed with a rate of 92.4%. This suggests that the generated text included question marks at the end and contained interrogative sentences. In contrast, Gemma-3-4B produced valid questions only 51.9% of the time. FLAN-T5-XL generated no valid questions, resulting in a 0% rate.
- Answer Leakage (no_answer_word): All models demonstrated high adherence to this constraint (>98%), ensuring that the answer key was not revealed within the question stem. Llama-3.2-3B achieved perfect compliance (100%).
- Explanation Prohibition (no_expl_word): Llama-3.2-3B was the most effective at excluding unwanted reasoning or explanations, achieving 67.1% compliance. Qwen2.5 followed with 60.8%. While FLAN-T5-XL shows 100% compliance here, this is an artifact of its failure to generate substantive text rather than successful instruction following.
- Length Constraints (length_ok): Qwen2.5-3B was the most successful at generating questions within the optimal length range (50.6%), whereas Llama-3.2-3B frequently generated output outside the target length parameters (10.1%)

E. Summary of Model Stability

Based on the "Output Stability" qualitative measure, Llama-3.2-3B and Qwen2.5-3B exhibited High stability, consistently following instructions and maintaining formatting. Mistral-7B and MediPhi showed Medium-to-High stability. Gemma-3-4B and FLAN-T5-XL were rated Low and Very Low, respectively, due to hallucinations and failure to adhere to the fundamental task of question generation. Qualitative samples confirm that while Llama and Qwen produced coherent, cognitively aligned questions (e.g., "Why did delegates think the Constitution needed limits on federal power?"), FLAN-T5-XL produced fragmented summaries.

To establish a solid basis for evaluation in educational measurement theory, a human assessment of cognitive alignment was performed on a random sample of 80 generated questions. Two human judges determined whether each question matched the corresponding Bloom level using a three-point correctness scale. The assessment yielded a Cohen's κ of 0.76, with 72% of the questions correctly aligned. This indicates that the improvements observed in the automatic metrics are genuine and reflect real enhancements in pedagogical validity. Statistical significance was assessed using paired t-tests on key metrics, confirming that the observed improvements are statistically significant ($p < 0.05$).

In an effort to determine how automatic metrics correlate with human assessment, a correlation analysis was performed using the evaluation metrics and human cognitive alignment assessment scores. The analysis shows a medium level of positive correlation between semantic similarity metrics, such as BERTScore and SBERT, and human assessment metrics. On the other hand, BLEU and ROUGE lexical metrics showed low correlation, owing to their focus on superficial similarities. This clearly shows that not all cognitive alignment can be captured by one evaluation metric. Instead, semantic metrics, combined with human assessment and validation using constraints, provide the best measures.

VII. Discussion

This study indicates that fine-tuned, instruction-aligned Large Language Models (LLMs) can reliably generate cognitively aligned educational assessment items when trained on enriched instructional data. By incorporating semantic similarity, lexical accuracy, and rule-based validation, this study provides an extensive understanding of model behaviour that goes beyond question generation. This section highlights the overall findings and their implications for educational assessment research.

First, the results clearly show a distinction between decoder-only and traditional encoder–decoder architectures. Specifically, Llama-3.2-3B and Qwen2.5-3B demonstrated superior performance compared to other models on both semantic preservation and output validity measures. Notably, Qwen2.5-3B achieved the highest F1 score on the BERTScore measure, indicating significant semantic preservation of the pedagogical intent encoded in the reference items. On the other hand, Mistral-7B achieved the highest score on the SBERT cosine similarity measure, indicating strong alignment at the sentence level.

Second, lexical-similarity metrics indicate that structural fidelity depends significantly on the model type. The best scores for the BLEU and METEOR metrics were achieved by the Llama-3.2-3B model, suggesting greater similarity in

TABLE V. PERFORMANCE COMPARISON OF FINE-TUNED LLMS

Category	Metric	Llama-3.2-3B	Qwen2.5-3B	Mistral-7B	MediPhi	Gemma-3-4B	FLAN-T5-XL
Semantic Similarity	BERTScore F1	0.8147	**0.8275**	0.8217	0.8068	0.447	0.7894
	SBERT Cosine	0.3029	0.3229	**0.3272**	0.3134	0.1909	0.0356
String-Matching Similarity	BLEU	**0.1131**	0.0607	0.0729	0.0677	0.0445	0
	ROUGE-L F1	0.069	0.0904	**0.0914**	0.0648	0.0096	0
	METEOR	**0.0989**	0.0718	0.0826	0.0758	0.0257	0
Question Validity	is_question	**94.90%**	91.10%	91.10%	92.40%	51.90%	0.00%
	length_ok	10.10%	**50.60%**	36.70%	27.80%	0.00%	0.00%
Constraint Compliance	no_answer_word	**100%**	98.70%	98.70%	98.70%	98.70%	100%
	no_expl_word	**67.10%**	60.80%	53.20%	55.70%	53.20%	100%

TABLE VI. QUALITATIVE COMPARISON OF GENERATED ASSESSMENT ITEMS ACROSS MODELS

Prompt (Shortened)	Ground Truth	Llama-3.2-3B	Qwen2.5-3B	Mistral-7B	MediPhi	Gemma-3-4B	FLAN-T5-XL
Humanities – Understand "Bill of Rights passage"	Why were the authors of the Federalist Papers against the Bill of Rights?	Why did some leaders at the Constitutional Convention believe a stronger federal government was necessary?	How did concerns about potential government tyranny influence the delegates' decision to add the Bill of Rights?	In what ways did fears about a powerful executive branch motivate the creation of the Bill of Rights?	Why were many delegates concerned that the new Constitution might threaten individual rights?	Why did delegates think the Constitution needed limits on federal power?	The need for rights in government.
Physics – Apply "Velocity graph passage"	How to calculate velocity and acceleration?	How would you determine the acceleration of an object using the slope of a velocity graph?	How can you calculate displacement using the area under a velocity graph?	How would you use slope and graph values to determine acceleration and displacement?	How do you apply the slope formula to find acceleration?	How can the slope of a velocity graph be used?	Using the graph to find acceleration.
Art History – Remember "Stonehenge passage"	(Example recall question)	What was the purpose of the Aubrey holes in the first phase of Stonehenge construction?	What significant feature characterized the first phase of Stonehenge construction?	When did the first phase of Stonehenge construction occur?	What was built during the first phase of Stonehenge?	What is Stonehenge?	Stonehenge construction phases.

phrasing and construction with human-generated texts. However, it is important to note that high lexical metric scores did not necessarily correlate with high cognitive alignment scores. This is also supported by some of Llama's outputs, which showed a tendency to paraphrase at the surface level rather than perform semantic-level transformations. This aligns with previous findings on LearningQ-based pipelines and indicates that it is not possible to generate higher-level questions solely through imitation of phrasing; instead, it requires cues and cognitive scaffolding incorporated into the instructional content. The relatively low performance on length constraints is due to the model's tendency to generate more descriptive questions.

Third, constraint evaluation for specific tasks is beneficial for understanding the feasibility of applying LLMs to standardized assessments. In this regard, metrics such as "is_question" and "no_answer_word" indicate that Llama-3.2-3B and MediPhi successfully constrained questions to maintain correct grammatical form and prevent answer words from appearing. In contrast, Qwen2.5-3B also scored highly on these metrics. Gemma-3-4B and FLAN-T5-XL, on the other hand, faced significant challenges with grammatical form and answer word inclusion. Moreover, while other models generated questions, FLAN-T5-XL failed to produce valid outputs. This suggests that model selection is critical when using AI to generate assessment questions. Finally, there is a critical observation regarding output stability, which is increasingly recognized as a primary requirement when applying LLMs to educational assessments. In this regard, Llama and Qwen scored highly on output stability. In contrast, while Mistral and MediPhi scored moderately, Gemma and FLAN-T5 scored very poorly. This is consistent with recent research suggesting that LoRA is most effective when applied to architectures explicitly pre-trained on instructional tasks.

Finally, the study used a data-enrichment pipeline. It included flattening the JSON structure of the raw data, using the Qwen2.5-7B model to generate cognitive labels, emphasizing key passage segments, and developing stringent templates for model outputs. The model's performance relied on this pipeline.

Despite these promising findings, there are also clear limitations: the experiments were conducted in a deterministic decoding setting, examined only models ranging from 3B to 7B in size, and used heuristic constraints that may not fully capture pedagogical quality. Overall, small- to medium-sized, instruction-tuned LLMs appear capable of generating cognitively aligned and structurally valid assessment items, putting models such as Llama and Qwen in a strong position for integration into various educational software solutions.

VIII. CONCLUSION AND FUTURE WORK

This study demonstrates the feasibility of generating cognitively aligned assessment questions efficiently using open-source, instruction-tuned language models, based on structured data augmentation. By incorporating evidence-based prompts and Bloom's taxonomy labels, the proposed framework improves both the structural validity and semantic alignment of the generated questions.

These results highlight the importance of selecting the appropriate architecture and designing effective prompts in

educational natural language processing systems, showing that decoder-only models are particularly well-suited for constrained generation tasks.

Future work should expand the scope of human evaluation of cognitive alignment by incorporating more detailed annotations and broader expert validation, developing bias-aware explanation strategies, and testing robustness under random decoding to further enhance educational reliability.

AUTHORS' CONTRIBUTIONS

Conceptualization, Mahmoud Badry and Shereen A. Taie; methodology, Mahmoud Badry; software, Mahmoud Badry; validation, Mahmoud Badry, Asmaa Sweidan, and Walaa Medhat; formal analysis, Mahmoud Badry; investigation, Mahmoud Badry; data curation, Mahmoud Badry; writing—original draft preparation, Mahmoud Badry and Asmaa Sweidan; writing—review and editing, Shereen A. Taie, Asmaa Sweidan, and Walaa Medhat; supervision, Shereen A. Taie and Asmaa Sweidan.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, TX, USA, 2016, pp. 2383–2392.

[2] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-Scale Reading Comprehension Dataset from Examinations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 785–794.

[3] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, "LearningQ: A Large-Scale Dataset for Educational Question Generation," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 12, no. 1, Stanford, CA, USA, Jun. 2018.

[4] Qwen Team, "Qwen2.5: A Party of Foundation Models," https://qwenlm.github.io/blog/qwen2.5/, 2024, qwen Blog.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.

[6] A. Dubey, "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.

[7] H. W. Chung, "Scaling Instruction-Finetuned Language Models," *arXiv preprint arXiv:2210.11416*, 2022.

[8] A. Hadifar, "EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain," *IEEE Access*, vol. 11, pp. 20 885–20 896, 2023.

[9] O. Holl, "EduQuest: Lecture Texts and Questions for Higher Education," in *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, Atlanta, GA, USA, 2024.

[10] S. Lamsiyah, "Fine-Tuning a Large Language Model with Reinforcement Learning for Educational Question Generation," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2024.

[11] M. A. Ehsan, A. S. Hasan, K. B. Shahnoor, and S. S. Tasneem, "Automatic Question and Answer Generation Using Generative Large Language Models (LLMs)," Ph.D. dissertation, Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh, 2024.

[12] Q. Zhuge, H. Wang, and X. Chen, "TwinStar: A Novel Design for Enhanced Test Question Generation Using Dual-LLM Engine," *Applied Sciences*, vol. 15, no. 6, p. 3055, 2025.

[13] P. Meyers, "Focal: A Proposed Method of Leveraging LLMs for Automating Assessments," in *Proceedings of the International Conference on Computers in Education (ICCE)*, 2023.

[14] S. García-Méndez, F. de Arriba-Pérez, and M. del C. Somoza-López, "A Review on the Use of Large Language Models as Virtual Tutors," *Science & Education*, vol. 34, no. 2, pp. 877–892, 2025.

[15] M. Stamatakis, "Enhancing the Learning Experience: Using Vision-Language Models to Generate Questions for Educational Videos," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2025.

[16] N. Scaria, S. D. Chenna, and D. Subramani, "Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2024.

[17] T. Huber and C. Niklaus, "LLMs Meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025.

[18] A. Q. Jiang, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[19] J.-P. Corbeil, "A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment," *arXiv preprint arXiv:2505.10717*, 2025.

[20] Gemma Team, "Gemma 3 Technical Report," *arXiv preprint arXiv:2503.19786*, 2025.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020, https://arxiv.org/abs/1904.09675.

[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992, https://arxiv.org/abs/1908.10084.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318, https://aclanthology.org/P02-1040/.

[24] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.

[25] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, 2005, https://aclanthology.org/W05-0909/.

A Machine Learning Approach for Targeting Lucrative Customer Segments

Leen Almajed[1], Haifa Almakhdhoub[2], Jana Alzeydan[3], Alaa Bin Sleem[4], Ouiem Bchir[5]
TAHAKOM, Riyadh, Saudi Arabia[1,2]
College of Computer and Information Sciences-Computer Science Department,
King Saud University, Riyadh, Saudi Arabia[1,2,3,4,5]

***Abstract*—Effective marketing plays a pivotal role in modern businesses, where strategic allocation of resources is essential for maximizing return on investment (ROI). Customer segmentation involves dividing users into distinct sub-groups based on common characteristics, enabling each segment to receive tailored promotions according to their behavior. However, identifying which customer segments to target can be financially challenging due to the significant costs associated with marketing campaigns. This study proposes a machine learning framework to forecast the profitability of various customer segments and optimize marketing strategies accordingly. The proposed solution leverages machine learning and deep learning techniques to classify customers based on their potential value. Specifically, conventional classifiers including AdaBoost, Gradient Boosting, Extreme Gradient Boosting (XGBoost), Linear Discriminant Analysis (LDA), Random Forest, Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are evaluated. In addition, deep learning models, namely 1D ResNet and 1D DenseNet, are investigated. All models are trained and evaluated under a unified protocol that includes SMOTE-based class imbalance handling, systematic hyperparameter tuning, and a joint analysis of predictive performance and computational cost. The experimental findings reveal that traditional models, particularly XGBoost and Gradient Boosting, consistently outperform deep learning models in terms of accuracy, precision, and computational efficiency, with both achieving the highest weighted F1 score of 0.62 while requiring nearly six orders of magnitude fewer computations than DenseNet-121. These results provide concrete evidence that ensemble tree-based methods are better suited than deep architectures for moderately sized, imbalanced, tabular marketing datasets.**

Keywords*—Return on investment; customer segmentation; machine learning; deep learning; profitability prediction; tabular data; Gradient Boosting; XGBoost

I. Introduction

As the business environment grows more competitive, companies increasingly leverage data-driven targeted marketing strategies to drive financial growth [1]. Targeted marketing identifies a segment of people with similar needs or characteristics that a company chooses to serve, allowing messages and promotions to speak directly to the interests of each segment rather than relying on a one-size-fits-all approach [2]. Customer segmentation that incorporates psychological and behavioral dimensions enables businesses to tailor their marketing efforts and increase customer retention by aligning campaigns with customer preferences [3]. This approach improves the relevance and effectiveness of marketing efforts, resulting in higher engagement and conversion rates. By focusing on segments with the highest growth potential or profitability, companies can prioritize their marketing budgets, product development initiatives, and sales efforts to maximize return on investment [2]. The integration of data analytics techniques, such as machine learning and predictive modeling, allows these strategies to be refined further to meet specific market needs [4].

Despite these advantages, many companies still face significant challenges in implementing targeted marketing effectively. In practice, some organizations continue to rely on mass or untargeted approaches, which can lead to substantial financial loss when resources are spent reaching individuals who are uninterested in the products on offer [5]. The result is a wasted budget and a diminished return on investment (ROI), since the lack of focus prevents companies from engaging the segments most likely to respond. Although customer segmentation partially addresses this issue, accurately determining the profitability of each segment under a given marketing strategy remains difficult. This difficulty arises from the complexity of predicting how different customer groups will respond to campaigns and how those responses translate into financial outcomes. Traditional customer targeting techniques often rely on simplistic criteria such as demographic categories or basic purchasing metrics, which may not capture the complex patterns within customer data that determine profitability. The interactions between customer attributes, for example how the combination of income level, purchase frequency, and geographic location jointly influences campaign response, are inherently non-linear and difficult to model using rule-based approaches. Consequently, there is a clear need for an intelligent, data-driven approach capable of forecasting the profitability of different customer segments.

To address these challenges, machine learning offers a promising solution by analyzing large-scale customer data to uncover patterns that support accurate profitability prediction. By examining demographic, geographic, behavioral, and psychographic attributes, machine learning models provide actionable insights that enable businesses to allocate resources toward high-value segments and optimize marketing strategies, ultimately enhancing ROI and long-term customer relationship management (CRM). The ability of machine learning to process high-dimensional data and identify non-obvious relationships makes it particularly valuable for profitability prediction, where the interplay between multiple customer

This work was conducted while Leen Almajed and Haifa Almakhdhoub were affiliated with the Computer Science Department, College of Computer and Information Sciences, King Saud University.

attributes determines segment-level outcomes. Furthermore, the availability of diverse algorithmic approaches, ranging from interpretable linear models to complex deep learning architectures, enables practitioners to select models that best match the characteristics of their specific datasets and business requirements.

Building on this approach, this study proposes a machine learning framework for predicting customer segment profitability using historical customer data and advanced classification techniques. The framework is designed to support strategic decision-making and efficient resource allocation. To evaluate performance and robustness, multiple conventional classifiers are compared, including AdaBoost [6], Gradient Boosting [7], XGBoost [8], LDA [9], Random Forest [10], Naïve Bayes [11], SVM [12], and KNN [13]. Additionally, 1D deep learning architectures, namely ResNet [14] and DenseNet [15], are assessed to determine their effectiveness in predicting segment-level profitability.

The main contributions of this work are as follows: 1) a unified empirical benchmark of eight classical classifiers and two 1D deep learning architectures (1D ResNet and 1D DenseNet) applied to customer segment profitability prediction on a previously unexplored, moderately sized, imbalanced tabular dataset; 2) a domain-specific contribution to the marketing analytics literature showing that, contrary to several recent reports favoring deep learning for customer prediction [16], [17], [18], traditional ensemble methods outperform 1D deep architectures when applied to comparative two-group customer features under realistic data-volume and class-imbalance conditions; 3) a detailed computational cost analysis that quantifies the trade-off between predictive performance and resource requirements across all evaluated models, providing the GFLOPs and parameter counts needed to assess deployment feasibility in production marketing systems; and 4) practical recommendations for practitioners on model selection, hyperparameter behavior under class imbalance, and the limited gains obtainable from autoencoder-based feature compression or One-vs-Rest decomposition on this class of data.

II. Related Work

A. Traditional Machine Learning

Several studies applied traditional machine learning methods to predict customer behavior. Deniz and Bülbül [19] applied six models, including Logistic Regression [20], Random Forest [10], and Gradient Boosting [7] to predict purchasing behavior using attributes like age, gender, annual income, and number of purchases. A significant improvement was achieved through segmenting customers into high, medium, and low spenders. Random Forest achieved the highest accuracy at 94.4%, outperforming Logistic Regression at 82.2%.

Other studies [21], [22] showed that Random Forest outperformed other models achieving 90.4% and 88.63% accuracy, respectively, when predicting customer subscription behaviors and churn by analyzing features like call duration, contact frequency, and customer behavioral metrics.

Extending the approach to customer profitability prediction, Chen et al. [23] introduced a dynamic approach using the Recency, Frequency, and Monetary (RFM) model. Customers were segmented into profitability groups using k-means clustering, and models including regression, multilayer perceptron (MLP), and Naïve Bayes were applied. Open-loop models achieved an average accuracy of 84%, while closed-loop models reached 79%. The study highlighted that customer profitability groups remained largely stable over time, with transition probabilities below 6%.

Saeed et al. [24] employed an ensemble classifier combining Random Forest and Extra Trees to predict customer responses to a banking telemarketing campaign. The hybrid voting classifier improved accuracy by 1.6% over the best individual model, Random Forest, which achieved 94.02%.

In related studies, the gradient boosting decision tree (GBDT) demonstrated strong performance. Sun et al. [25] applied GBDT and Random Forest to predict Customer Lifetime Value (CLV), achieving mean square errors of 0.65 and 0.34, respectively, outperforming probability-based models. Similarly, Sun et al. [26] compared eight machine learning algorithms for customer segmentation based on lifecycle value, where AdaBoost achieved 94.60% accuracy and GBDT reached 93.80%.

Xiahou and Harada [27] utilized AdaBoost for churn prediction, evaluating it against a Backpropagation (BP) neural network. The AdaBoost model outperformed the BP neural network across all metrics, achieving accuracy, recall, and precision values of 0.96, 0.93, and 0.96, respectively. Mishra and Sharma [28] further validated AdaBoost for fraud detection, showing its ability to handle imbalanced datasets effectively.

Wong and Yeh [29] proposed hybrid classification algorithms that combine instance filtering with ensemble methods, demonstrating improved accuracy over standalone classifiers on customer behavior datasets. Uddin et al. [30] compared multiple supervised algorithms, including SVM, KNN, and Random Forest, finding that ensemble methods consistently outperformed individual classifiers for structured prediction tasks. Musa [31] conducted a comparative study between SVM and Logistic Regression, concluding that SVM achieves superior classification performance on high-dimensional customer datasets.

B. Deep Learning

Deep learning models have been increasingly adopted for customer prediction tasks. Xu et al. [16] employed a combination of ResNet-50 [14] and Gated Recurrent Unit (GRU) [32] to predict consumer behavior by combining visual and temporal data, achieving a prediction accuracy of 96.18%.

Keji et al. [17] applied a ResNet model to banking data [33] with Synthetic Minority Oversampling Technique (SMOTE) for class imbalance. ResNet achieved 93% accuracy, outperforming Random Forest at 90.78%. Nabi et al. [18] focused on profit prediction in e-commerce using deep learning. VGG16 [34] and ResNet50 [14] were compared to traditional machine learning algorithms. VGG16 emerged as the top-performing model with 92.55% accuracy.

Chaudhuri et al. [35] compared a Feed-forward Deep Neural Network (DNN) to traditional models for predicting customer purchase behavior on e-commerce platforms. DNN

outperformed all traditional algorithms with an accuracy of 89%. Singh et al. [36] compared MLP [37] to decision tree J48 and Naïve Bayes classifiers on a shopping center dataset for Customer Relationship Management (CRM). MLP outperformed both classifiers with an accuracy of 98.33%.

Dong et al. [38] proposed an improved 1D ResNet architecture for structured data classification, demonstrating that residual connections are effective even in non-image domains. Wang et al. [39] established that deep neural networks, including ResNet and fully convolutional networks provide strong baselines for time series and structured data classification, often matching or exceeding traditional approaches without extensive feature engineering.

While deep learning models excel with multi-modal and unstructured data, their application to structured tabular data remains less common. Traditional models are often preferred for their interpretability and suitability for smaller structured datasets. Recent benchmark studies have shown that gradient boosting methods frequently outperform deep learning on tabular data, particularly when datasets contain fewer than 10,000 instances. This study uses a dataset not previously explored in the literature and focuses on an empirical comparison of both paradigms for customer segment profitability prediction, providing evidence for the relative strengths of each approach under conditions of moderate dataset size and class imbalance.

III. Proposed Approach

This study proposes a machine learning framework for customer profitability prediction. The input to this framework is segmented historical customer data, structured based on customer attributes. This data undergoes a preprocessing pipeline that includes: 1) feature standardization to normalize input variables, ensuring that features with different scales do not disproportionately influence model training, 2) application of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the training set by generating synthetic instances of the minority class, and 3) for deep learning models, reshaping of input features to match the required format for 1D convolutional architectures. SMOTE is applied only to the training set to prevent data leakage, ensuring that the validation and test sets reflect the true class distribution.

The framework follows two main pathways. The first pathway trains and evaluates conventional machine learning models, including AdaBoost [6], Gradient Boosting [7], Extreme Gradient Boosting [8], Linear Discriminant Analysis [9], Random Forest [10], Naïve Bayes [11], Support Vector Machines [12], and K-Nearest Neighbors [13]. These models are trained using 10-fold cross-validation, where the data is partitioned into ten equal subsets, and each model is trained on nine folds while being validated on the remaining fold, rotating across all ten combinations. This strategy provides a robust estimate of model performance by reducing the variance associated with a single train-test split. The second pathway trains deep learning models, specifically 1D ResNet [14] and 1D DenseNet [15]. For deep learning, the dataset is split into 60% training, 20% validation, and 20% testing, with early stopping employed to halt training when validation performance ceases to improve, thereby preventing overfitting.

Following model training and testing, each model undergoes evaluation to assess performance. Based on the evaluation results, a final decision is made to select the best-performing model. The chosen model is then used to predict the profitability of each customer segment, providing actionable insights for targeting high-value customer segments and optimizing marketing resources.

The motivation for including conventional ML models stems from their advantages with structured datasets: they are computationally less intensive, easier to interpret, and generally less prone to overfitting with smaller datasets. Past research consistently demonstrates their effectiveness in customer-related prediction tasks [19], [21], [25], [27]. Additionally, conventional models allow for quicker training times, and their results are more interpretable, which is beneficial when presenting findings to stakeholders unfamiliar with deep learning concepts.

Deep learning models, specifically 1D ResNet and 1D DenseNet, are included due to their sophisticated feature extraction capabilities. ResNet uses residual learning with skip connections that alleviate the vanishing gradient problem [40], enabling more intricate pattern learning. In the context of customer data, this can mean better learning of hierarchical feature interactions across various attributes. DenseNet connects each layer to every subsequent layer in a feed-forward fashion, improving information flow and reducing redundant feature learning. This dense connectivity pattern allows DenseNet to retain critical customer information across layers, leading to more accurate predictions with fewer parameters compared to equivalently deep standard architectures. The choice of 1D architectures aligns with the structured nature of customer data, where each record represents a customer attribute vector, enabling the models to process tabular features with minimal transformations.

IV. Experiments

A. Dataset Description

The dataset [41], "Predicting Profitable Customer Segments", is a publicly available Kaggle benchmark that aggregates historical records from an online retail company to support predictive modeling for targeted marketing campaigns. Each row represents a pre-campaign observation that compares two customer groups, capturing their attributes and relational characteristics before any promotional intervention.

The dataset contains 6621 instances and 70 input features, organized into three semantically distinct categories of columns: the "g1_" columns describing absolute attributes of the first customer group (e.g., aggregated demographic and behavioral measurements), the "g2_" columns describing the corresponding attributes of the second group, and the "c_" columns capturing relational features that quantify pairwise differences between the two groups. The target column categorizes outcomes into three classes: neither group was profitable (0), Group 1 was more profitable (1), or Group 2 was more profitable (2). The class distribution is moderately imbalanced: class 1 (Group 1 more profitable) is the majority class with 3076 instances (46.5%), followed by class 2 (Group 2 more profitable) with 1877 instances (28.3%), and class 0 (neither profitable) as the minority class with 1667 instances

(25.2%). Although exact attribute semantics are not disclosed by the data provider for confidentiality reasons, the variable names and value ranges indicate that the features encompass demographic descriptors (such as age, income, and geographic indicators) and behavioral measures (such as transaction value and purchase frequency). The comparative "c_" columns encode the relative positioning of the two groups, for example the difference in average spending or the ratio of engagement frequencies between the groups. This dual representation, combining absolute group characteristics with relative inter-group metrics, creates a feature space where models can learn both what makes individual groups profitable and how the contrast between groups predicts campaign outcomes. The class imbalance poses an additional challenge, as classifiers tend to optimize for majority classes unless explicitly guided otherwise through class weighting or oversampling. We acknowledge that the limited public documentation of feature semantics constrains the depth of feature-level interpretation that can be performed, and we discuss the implications of this limitation in Section VI.

The implementation used Google Colab Pro with TPU support, TensorFlow/Keras [42], [43] for deep learning, Scikit-learn [44] for classical models, and the XGBoost library [8]. NumPy [45] was used for data manipulation and Matplotlib [46] for visualization. Initial experiments were conducted on local machines equipped with NVIDIA GeForce GTX 1660 Ti and Apple M1 Pro GPUs. However, the computational demands of training deep architectures such as 1D ResNet-50 and 1D DenseNet-121 required migrating to Google Colab Pro, which provided access to TPU and high-RAM environments that significantly accelerated training.

B. Experiment 1: Deep Learning

Two deep learning architectures were evaluated: 1D ResNet and 1D DenseNet. All models were trained using a consistent preprocessing pipeline including feature standardization, SMOTE for class imbalance, and input reshaping to conform to the expected input format of 1D convolutional networks. The dataset was split into 60% training, 20% validation, and 20% testing, with early stopping to prevent overfitting. Running these experiments introduced significant computational overhead, as deep learning architectures such as 1D ResNet-50 and 1D DenseNet-121 were inherently expensive to train in terms of both time and memory, requiring access to TPU-accelerated cloud environments.

1) 1D ResNet: Three variants (ResNet-18, ResNet-34, ResNet-50) were evaluated. The Adam optimizer was used with a batch size of 128 and early stopping. Class weighting was applied to address imbalanced learning. Multiple configurations were explored by varying the learning rate from 0.0001 to 0.1 across four settings per variant, yielding 12 configurations total. The number of training epochs ranged from 21 to 44 due to early stopping.

Among all variants and configurations, ResNet-18 with learning rate 0.1 achieved the best performance, as shown in Table I. The small decline between validation and test performance indicates good generalization. ResNet-34 and ResNet-50 peaked at validation F1 scores of 0.496 and 0.487, respectively, both with learning rate 0.01, falling short of ResNet-18's 0.515. The finding that deeper variants did not improve over ResNet-18 can be attributed to the dataset's relatively small size and low complexity, where a simpler architecture with fewer parameters is more efficient at capturing essential patterns without memorizing noise. The higher learning rate of 0.1 outperformed smaller rates across all variants, suggesting that the loss landscape for this task is relatively smooth, allowing aggressive gradient updates to converge faster without oscillating past optima.

TABLE I. RESNET-18 PERFORMANCE WITH BEST CONFIGURATION

	Precision	Recall	F1 Score	Accuracy
Validation	0.5119	0.5431	0.515	0.5431
Test	0.4971	0.505	0.501	0.4562

TABLE II. DENSENET-121 PERFORMANCE WITH BEST CONFIGURATION

	Precision	Recall	F1 Score	Accuracy
Validation	0.54	0.55	0.55	0.55
Test	0.54	0.54	0.54	0.54

2) 1D DenseNet: DenseNet-121 was evaluated across 11 configurations varying the optimizer (Adam, AdamW), learning rate (0.001 to 0.1), and class weighting strategies (none, automatically computed based on class distribution, and manually assigned). The batch size was fixed at 128 with early stopping.

The experiments revealed critical insights about training deep models on imbalanced data. Configurations without class weights achieved moderate accuracy but ignored the minority class 0 entirely, resulting in low F1 scores. For instance, configurations with no class weights produced biased predictions that favored majority classes, achieving an F1-score of only 0.47. Configurations with very low learning rates (0.001) performed poorly regardless of class weighting, with F1 as low as 0.30, suggesting the optimizer converged too slowly or became stuck in shallow local minima. In contrast, configurations with appropriately set class weights and higher learning rates achieved substantially better results. These findings demonstrate that both an appropriate learning rate and effective class weighting are jointly necessary to handle imbalanced datasets; neither alone is sufficient. The AdamW optimizer with automatic class weights achieved competitive F1 of 0.54, but did not surpass the Adam optimizer under optimal settings. Automatic class weights, computed proportionally from the class distribution, provided a reasonable approximation but were less effective than manually assigned weights that doubled the penalty for misclassifying the minority class.

The best performance was achieved with Adam optimizer, learning rate of 0.1, and manually assigned class weights $\{0 : 2.0, 1 : 1.0, 2 : 1.0\}$, which doubled the penalty for misclassifying the minority class, as shown in Table II. The consistent results across validation and test sets indicate good generalization and no overfitting.

Overall, DenseNet-121 outperformed all ResNet variants, achieving a test F1 of 0.54 compared to 0.50 for ResNet-18. DenseNet benefits from its dense connectivity pattern that creates shorter paths for gradient flow and enables more efficient feature reuse across layers. Its concatenation-based approach

TABLE III. NAÏVE BAYES PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.50	0.49	0.49	0.48
Test	0.49	0.48	0.49	0.48

retains feature maps from all previous layers, allowing the model to access both low-level and high-level representations simultaneously. This is particularly beneficial for tabular data where feature interactions may span over different levels of abstraction. Deeper ResNet variants (ResNet-34, ResNet-50) did not improve over ResNet-18, suggesting that model depth should be proportional to the complexity and volume of the data. ResNet-50 achieved the lowest validation F1 among all variants (0.356 at learning rate 0.0001), indicating that the additional depth introduced unnecessary parameters that the limited training data could not support. The consistent finding that shallower architectures outperform deeper ones across both ResNet and DenseNet experiments suggests that the dataset's 6621 instances are insufficient to train deep networks effectively, and that the feature space does not contain hierarchical patterns complex enough to justify deep architectures.

C. Experiment 2: Traditional Machine Learning

Eight conventional machine learning classifiers were evaluated using 10-fold cross-validation during training. All models were trained on the same preprocessed dataset using feature standardization and SMOTE for class imbalance. For each model, multiple hyperparameter configurations were explored, and only the best-performing configuration is reported. The hyperparameters for each model were tuned systematically, varying key parameters such as learning rate, regularization strength, kernel type, and ensemble size to identify the optimal balance between model complexity and generalization.

1) Naïve Bayes: Naïve Bayes has no significant tunable hyperparameters and was evaluated using 10-fold cross-validation. Table III reports the results. The model shows consistent but low performance across training and test sets, with F1 below 0.5 reflecting its limited capacity on this dataset. This weak performance is likely due to the strong feature independence assumption, which rarely holds in real-world datasets where features tend to be correlated. The relational features in this dataset ("c_" columns quantifying group differences) are inherently derived from the group-specific features, creating strong inter-feature dependencies that directly violate this assumption. Despite this, the model's stability between training and test sets confirms that the issue is underfitting rather than overfitting. The near-identical training and test scores (0.49 vs. 0.49 for F1) indicate that the model has fully learned what its limited architecture permits, and additional data or features would not improve its performance without relaxing the independence assumption. While Naïve Bayes is the most lightweight model in our study, its weak predictive power limits its practical utility for this task.

2) K-Nearest Neighbors: KNN was tested with K values of 1, 2, 3, 5, 7, and 10 using 10-fold cross-validation. K=1 resulted in severe overfitting (training F1=1.0, test F1=0.71), as the model memorized the training data by returning the label of the single closest point. K=2 and K=3 achieved identical test F1 of 0.62 despite different training scores (0.84 and 0.80, respectively), highlighting diminishing returns from increased model complexity. K=5 yielded test F1 of 0.59, while K=7 dropped to 0.55. The generalization gap narrowed progressively from 0.29 (K=1) to 0.07 (K=10), illustrating the classic bias-variance tradeoff: lower K values produce models with low bias but high variance that are sensitive to noise, whereas higher K values produce smoother decision boundaries with higher bias but lower variance. K=10 provided the most balanced behavior, as shown in Table IV. Although its absolute test F1 of 0.52 is lower than K=2 and K=3, the reduced overfitting and more stable predictions make it a more reliable choice, particularly when the training set has been augmented with SMOTE-generated synthetic samples. The sensitivity of KNN to the choice of K underscores the importance of hyperparameter tuning and cross-validation in preventing misleading performance estimates.

TABLE IV. KNN PERFORMANCE RESULTS (K=10)

	Precision	Recall	F1 Score	Accuracy
Training	0.63	0.60	0.59	0.60
Test	0.56	0.53	0.52	0.53

TABLE V. RANDOM FOREST PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.55	0.56	0.54	0.56
Test	0.53	0.54	0.53	0.55

3) Random Forest: Three configurations were evaluated varying max depth (5–10), max features (sqrt, log2), and min samples parameters with 5 trees. All configurations achieved similar test performance (F1 between 0.53–0.54), indicating that Random Forest is relatively robust to hyperparameter changes on this dataset. The first two configurations with max depth 10 both achieved training F1 of 0.56 with test F1 of 0.53, exhibiting slight overfitting as the deeper trees memorized training patterns that did not generalize. Using sqrt versus log2 for max features had negligible impact, suggesting that feature diversity among trees was sufficiently maintained by either approach. The most constrained configuration with max depth 5 and increased min samples split provided the best generalization, as shown in Table V. This suggests that on this dataset, controlling tree complexity is more beneficial than allowing the model to capture finer-grained patterns that may be noise. The relatively small ensemble size of 5 trees was chosen to reduce training time, though increasing this number could potentially improve performance at the cost of additional computation.

4) Support Vector Machine: Over 120 hyperparameter configurations were explored, varying the kernel type (linear, polynomial, sigmoid, RBF), regularization parameter C, and gamma. All top-performing configurations used the RBF kernel, indicating its effectiveness in modeling the complex, non-linear decision boundaries present in this dataset. The RBF kernel maps data into a higher-dimensional feature space using a Gaussian function, enabling the construction of non-linear decision boundaries that are necessary for datasets with over-

TABLE VI. SVM PERFORMANCE RESULTS (RBF, C=0.5, GAMMA=0.01)

	Precision	Recall	F1 Score	Accuracy
Training	0.58	0.58	0.58	0.58
Test	0.56	0.53	0.54	0.53

TABLE VII. ADABOOST PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.55	0.56	0.55	0.56
Test	0.54	0.55	0.54	0.55

TABLE VIII. GRADIENT BOOSTING PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.70	0.70	0.69	0.70
Test	0.63	0.62	0.62	0.62

TABLE IX. XGBOOST PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.66	0.68	0.68	0.68
Test	0.60	0.61	0.62	0.62

lapping class distributions. A clear trend emerged: performance declined as C increased and gamma decreased, suggesting that overly large regularization values combined with very small gamma values harmed generalization. Specifically, the top six configurations all used RBF with C values ranging from 0.5 to 500 and gamma from 0.01 to 0.0001, with training F1 scores declining from 0.582 to 0.531 as C increased. The RBF kernel with C=0.5 and gamma=0.01 yielded the best results, as reported in Table VI. The moderate C value strikes a balance between maintaining a strict decision boundary and allowing flexibility for noisy data points. The training-test gap of 0.04 in F1 indicates moderate generalization. The linear, polynomial, and sigmoid kernels all performed below the RBF configurations, confirming that the class boundaries require the flexibility of a radial basis function.

5) AdaBoost: Three configurations were tested varying the learning rate (0.01, 0.1, 1.0) with 100 estimators and the SAMME algorithm [47]. A learning rate of 0.01 severely underfit with F1 of only 0.42 on both training and test sets, indicating that each weak learner's contribution was too small for the ensemble to converge within 100 iterations. This extreme consistency between training and test performance at a low score is a hallmark of underfitting, where the model has insufficient capacity to learn the underlying patterns. Learning rates of 0.1 and 1.0 achieved comparable test F1 of 0.54, with the higher learning rate (1.0) producing marginally higher training performance (0.55 vs. 0.54) without improving test results. The best result, shown in Table VII, demonstrates a minimal gap between training F1 (0.55) and test F1 (0.54), indicating no overfitting. This consistency highlights AdaBoost's strength in reducing variance through its sequential ensemble of weak learners, which iteratively increases the weights of misclassified instances to focus subsequent learners on harder cases. However, its absolute performance falls short of the more advanced boosting variants (Gradient Boosting and XGBoost), which employ gradient-based optimization of the loss function rather than AdaBoost's sample reweighting scheme, representing a 15% relative improvement in F1.

6) Gradient Boosting: Four configurations were tested varying the number of estimators (8–100), max depth (2–3), learning rate (0.1–1.0), and subsample ratio, all using Friedman-MSE criterion and log-loss. Configurations with fewer estimators or shallower trees showed progressively lower performance, indicating underfitting. The second configuration (10 estimators, max depth 3, learning rate 1.0) achieved test F1 of 0.59, while the third and fourth configurations (10 and 8 estimators, max depth 2) yielded 0.57 and 0.56, respectively. The best configuration used 100 estimators, max depth 3, and learning rate 0.1. As shown in Table VIII, Gradient Boosting achieved a test F1 score of 0.62, the highest among all models, tied with XGBoost. The gap between training F1 (0.69) and test F1 (0.62) is moderate, indicating that while some overfitting exists, the model still generalizes reasonably well. The combination of a moderate learning rate with sufficient estimators allowed the model to learn incrementally and capture subtle patterns, while max depth 3 prevented individual trees from becoming overly complex. The progressive performance improvement from the weakest to the strongest configuration demonstrates that the number of estimators and learning rate are the most influential hyperparameters for this task, with max depth playing a secondary role in controlling overfitting.

7) Extreme Gradient Boosting (XGBoost): Four configurations were evaluated varying max depth (2–4), learning rate (0.2–0.3), subsample ratio, and L1/L2 regularization. The best configuration used 100 estimators, max depth 4, learning rate 0.3, and reg_lambda 800 for strong L2 regularization. The high regularization parameter was critical in preventing overfitting despite the relatively deep trees, effectively penalizing large weights and keeping the model from becoming overly complex. The colsample_bytree of 0.8 introduced randomness by selecting a subset of features per tree, and the min_child_weight of 5 ensured that splits only occurred when sufficient data supported them. As shown in Table IX, XGBoost matched Gradient Boosting with a test F1 score of 0.62. Notably, a second configuration with only 50 estimators also achieved test F1 of 0.62 but with lower training performance (0.64 vs. 0.68), indicating that XGBoost can achieve comparable results with different hyperparameter combinations due to the compensatory relationships among its regularization parameters. The remaining two configurations with max depth 2 achieved lower test F1 scores of 0.57 and 0.59, confirming that deeper trees were necessary to capture the complexity of the dataset. The training-test gap of 0.06 is smaller than Gradient Boosting's gap of 0.07, suggesting slightly better generalization. XGBoost's built-in L1 and L2 regularization, combined with its optimized parallel processing and efficient handling of missing data, contributed to its strong performance while maintaining computational efficiency.

8) Linear Discriminant Analysis: Five configurations were tested varying the solver (SVD, LSQR, Eigen), shrinkage, prior probabilities, and tolerance. The SVD solver served as a baseline due to its computational efficiency and numerical stability, while the LSQR and Eigen solvers supported shrinkage to address potential feature redundancy. The Eigen solver

TABLE X. LDA PERFORMANCE RESULTS

	Precision	Recall	F1 Score	Accuracy
Training	0.54	0.53	0.53	0.53
Test	0.56	0.51	0.53	0.51

with automatic shrinkage yielded the best results, as shown in Table X. This combination provided the most balanced results across training and test sets, with the automatic shrinkage estimating an appropriate regularization level from the data. The SVD-based configurations, including one with uniform class priors, showed slightly lower performance, while increasing the tolerance to 0.5 produced the weakest results, suggesting that discarding too many dimensions reduced the model's discriminative ability. Despite testing multiple configurations, the performance differences across all LDA settings were relatively small (F1 ranging from 0.523 to 0.533), suggesting that LDA's linear projection is inherently limited for this dataset. The three-class problem with overlapping class distributions likely requires non-linear decision boundaries that LDA cannot model, which explains its lower performance compared to tree-based and kernel-based methods.

D. Additional Experiments

A One-vs-Rest (OvR) strategy was explored to improve classification of class 0, which the confusion matrix revealed was frequently misclassified as class 1 or class 2. This misclassification pattern was consistent across most models, indicating that class 0 (neither group profitable) shares significant feature overlap with both class 1 and class 2. Separate binary classifiers were trained to distinguish each class from the others, with a specific focus on isolating class 0. The rationale was that decomposing the multi-class problem into simpler binary tasks might allow classifiers to better learn the subtle differences between the minority class and the majority classes. However, the best F1 for class 0 reached only 0.4987, with other configurations yielding 0.4769, 0.4587, and 0.4335. These results suggest that the underlying class overlap and imbalance were difficult to overcome even with targeted binary modeling, and that the features available in the dataset may not be sufficiently discriminative for class 0.

Autoencoders [48], [49] were also tested as a preprocessing step to extract more meaningful features before classification. The hypothesis was that unsupervised feature learning could discover a compressed representation that better separates the classes. A shallow architecture (one hidden layer, 1000 neurons) achieved weighted F1 of 0.5245. However, increasing depth consistently degraded performance: a two-layer architecture yielded 0.5170, a three-layer version dropped to 0.5023, and further increasing the number of hidden nodes to 1500 also did not help, yielding 0.5023. Additional depth beyond three layers produced even lower scores of 0.4982, while reducing the number of nodes led to 0.4790. These results suggest that the dataset may not benefit from unsupervised compression, and that the original feature space already provides a sufficiently informative representation. The degradation with deeper architectures mirrors the findings from the ResNet experiments, reinforcing that additional model complexity does not improve results on this dataset.

TABLE XI. TEST PERFORMANCE COMPARISON OF ALL MODELS

Model	Prec.	Rec.	F1	Acc.	AUC
XGBoost	0.60	0.61	0.62	0.62	0.75
Gradient Boosting	0.63	0.62	0.62	0.62	0.74
AdaBoost	0.54	0.55	0.54	0.55	0.74
Random Forest	0.53	0.54	0.53	0.55	0.73
1D DenseNet-121	0.54	0.54	0.54	0.54	0.72
LDA	0.56	0.51	0.53	0.51	0.72
SVM (RBF)	0.56	0.53	0.54	0.53	0.66
KNN (K=10)	0.56	0.53	0.52	0.53	0.65
Naïve Bayes	0.49	0.48	0.49	0.48	0.69
1D ResNet-18	0.50	0.51	0.50	0.46	0.64

V. COMPARISON OF APPROACHES

Table XI summarizes the test performance of all evaluated models. XGBoost achieved the highest AUC of 0.75, followed by Gradient Boosting (0.74) and AdaBoost (0.74). Among deep learning models, DenseNet-121 performed best with an AUC of 0.72, while ResNet-18 reached 0.64.

Traditional machine learning models, particularly XGBoost and Gradient Boosting, consistently outperformed deep learning approaches. Both achieved the highest test F1 score of 0.62. XGBoost reached the highest AUC of 0.75, while Gradient Boosting followed closely at 0.74. Three interacting properties of the dataset jointly explain this gap. First, the data is tabular and heterogeneous: each feature has its own semantics and scale, so the translation-invariant local filters learned by 1D convolutional architectures have no clear analogue in this representation, whereas axis-aligned splits used by tree-based ensembles map directly onto the per-feature decision structure that is natural for this type of data. Second, the dataset is moderately sized: deep architectures typically require large amounts of training data to generalize effectively, and the 60/20/20 split further reduces the effective deep learning training set to approximately 3973 instances, whereas traditional models benefit from 10-fold cross-validation that uses 90% of the data for training in each fold. Third, the class distribution is imbalanced, and although SMOTE and class weighting mitigate this, the residual difficulty of separating the minority class 0 from classes 1 and 2 affects high-capacity deep models more severely than tree ensembles, which can isolate small but informative regions of feature space through targeted splits. These three factors, namely tabular structure, moderate sample size, and class imbalance, are precisely the conditions under which gradient boosting has been repeatedly shown to dominate deep learning on tabular benchmarks, and our results corroborate that pattern in the marketing analytics setting.

A consistent observation across all models is that the weighted F1 score appears to be bounded around 0.62. This bound is not a tuning artefact: it persists across systematic hyperparameter searches (over 120 configurations for SVM, multiple configurations for every other model), across an alternative classification strategy (One-vs-Rest, with the best class-0 F1 reaching only 0.4987), and across an alternative feature representation (autoencoder-based compression, peaking at 0.5245). The fact that two architecturally distinct top models (XGBoost and Gradient Boosting) converge to the same 0.62 score, and that neither deeper trees, nor stronger regularization, nor unsupervised feature extraction can push

past it, indicates that the limit is a property of the data rather than of the classifiers. The confusion-matrix analysis (see Section IV) confirms that the residual errors concentrate on class 0, which the feature space cannot reliably distinguish from the other two classes given the available attributes. This is consistent with the dataset's confidentiality-induced opacity: domain-relevant features that might disambiguate class 0 (such as marketing channel, prior campaign exposure, or product-category preference) may simply not be represented in the published feature set. Improving beyond F1 $\approx$ 0.62 would therefore most likely require richer input features rather than more expressive models, which we discuss as a direction for future work.

Among deep learning models, DenseNet-121 outperformed ResNet-18, achieving a test F1 score of 0.54 compared to 0.50. ResNet-18 outperformed deeper ResNet variants (ResNet-34 and ResNet-50), suggesting that a simpler architecture was better suited for this dataset. Notably, DenseNet-121 achieved performance comparable to several traditional models (AdaBoost at 0.54, SVM at 0.54) despite requiring orders of magnitude more computation, which raises questions about the practical value of deep learning for structured datasets of this scale.

An interesting observation is the relationship between AUC and F1 discrepancies: Naïve Bayes achieved AUC of 0.69 despite an F1 of only 0.49, indicating that while it can partially rank instances by class probability, its hard classification decisions are poor. This suggests that Naïve Bayes produces well-calibrated probability estimates despite its inability to set effective classification thresholds. LDA similarly showed a gap between AUC (0.72) and F1 (0.53), reflecting its strength in linear separability assessment despite limited classification accuracy. Conversely, XGBoost achieved the highest AUC alongside the highest F1, demonstrating both strong ranking ability and effective classification thresholds. This consistency between ranking and classification metrics indicates that XGBoost produces reliable probability estimates that translate directly into accurate predictions.

Several models incorporate implicit feature selection mechanisms. XGBoost, Random Forest, and AdaBoost evaluate feature importance during tree construction, naturally prioritizing informative features. LDA performs dimensionality reduction by projecting data onto a space that maximizes class separability. SVM with RBF kernel maps data into higher-dimensional spaces for better separation. Deep learning models rely on convolutional layers to learn hierarchical feature representations automatically. These built-in mechanisms allow each model to handle high-dimensional inputs differently, contributing to the observed performance differences.

In terms of computational efficiency, traditional models were significantly more resource-efficient. Table XII and Table XIII compare the computational requirements of all models. The contrast is substantial: DenseNet-121 required 121232 GFLOPs for training versus only 0.15 GFLOPs for XGBoost, a difference of nearly six orders of magnitude. In parameter count, DenseNet-121 (5.53M) had over 120 times more parameters than XGBoost (0.04M). Even the lighter deep learning model, ResNet-18 (0.23M parameters, 3582 GFLOPs), demanded over 24000 times more computation than XGBoost.

TABLE XII. SPACE COMPLEXITY (NUMBER OF PARAMETERS)

Model	Params (M)
1D DenseNet-121	5.5328
SVM	0.4856
KNN	0.4635
1D ResNet-18	0.2342
XGBoost	0.0445
Gradient Boosting	0.0213
LDA	0.0027
AdaBoost	0.0015
Random Forest	0.0004
Naïve Bayes	0.0002

TABLE XIII. TIME COMPLEXITY (GFLOPS)

Model	Training	Testing
1D DenseNet-121	121231.62	10099.90
1D ResNet-18	3582.26	303.06
SVM	1.96	0.10
Gradient Boosting	0.46	0.01
XGBoost	0.15	0.001
LDA	0.03	0.0001
AdaBoost	0.02	0.005
Random Forest	0.002	$<$0.001
Naïve Bayes	$<$0.001	0.001

Naïve Bayes was the most lightweight model (0.0002M parameters, $<$0.001 GFLOPs), but its weak predictive performance limits its practical utility. KNN, while requiring no formal training phase, incurs inference-time cost proportional to the dataset size (0.49 GFLOPs for testing), storing all 0.46M training instances for distance computation at prediction time. SVM, despite achieving moderate F1 of 0.54, required 1.96 GFLOPs for training due to the kernel computations needed for constructing the decision boundary in high-dimensional space. Among the top-performing models, XGBoost was notably faster than Gradient Boosting (0.15 vs. 0.46 GFLOPs for training) due to its optimized parallel processing and built-in regularization, while both achieved identical test F1 of 0.62. XGBoost struck a favorable balance with 0.04M parameters and strong predictive performance.

The results also reveal a clear performance hierarchy within the traditional ML models. The boosting family (XGBoost, Gradient Boosting, AdaBoost) consistently outperformed other approaches, with the gradient-based variants (XGBoost and Gradient Boosting) achieving the best results. This advantage stems from their iterative error correction mechanism, where each subsequent tree focuses on the residuals of the previous ensemble. Random Forest, despite being an ensemble method, achieved lower F1 (0.53) due to its independent tree construction that does not explicitly target misclassified instances. SVM and KNN achieved moderate performance (F1 of 0.54 and 0.52, respectively), while Naïve Bayes and LDA, both limited by strong distributional assumptions, fell short. This ordering aligns with the general principle that models capable of capturing non-linear interactions and iteratively refining predictions perform best on complex classification tasks.

Overall, XGBoost and Gradient Boosting demonstrated the best balance between accuracy, generalization, and computational efficiency, making them well-suited for the customer segment profitability prediction task. Their suitability for structured tabular data and their interpretability provide a practical

advantage for business decision-making, where understanding which customer group features drive profitability predictions is as valuable as the predictions themselves.

From a deployment perspective, the computational disparity reported in Table XII and Table XIII translates directly into operational consequences. XGBoost requires only 0.001 GFLOPs at inference time compared with 10099.90 GFLOPs for DenseNet-121, making it roughly ten million times faster at prediction time and enabling a profitability score to be produced for an incoming customer segment in essentially constant time on commodity CPU hardware. This makes XGBoost compatible with real-time campaign-optimization pipelines and with edge or in-database deployment. By contrast, DenseNet-121 effectively assumes GPU/TPU availability and introduces additional infrastructure cost, latency, and energy consumption that are hard to justify given the absence of any predictive advantage. The 0.04M parameter footprint of XGBoost also simplifies model versioning, A/B testing, and incremental retraining, which are routine requirements in marketing analytics platforms. Finally, the tree-based nature of XGBoost preserves direct compatibility with established interpretability tools (such as gain-based feature importance and SHAP values), supporting the auditability and explanation requirements that increasingly accompany automated marketing decisions. Taken together, these properties make XGBoost and Gradient Boosting the recommended candidates for production deployment in this domain, with deep architectures reserved for settings where dataset size and modality (for instance, image- or sequence-rich customer data) genuinely warrant them.

VI. Limitations

Several limitations of this study should be acknowledged explicitly. First, the dataset size of 6621 instances is moderate relative to the data volumes at which 1D ResNet and 1D DenseNet have historically demonstrated their advantages, and a fairer assessment of deep learning would require benchmarks with at least an order of magnitude more samples; our negative result for deep learning should, therefore, be interpreted as conditional on the data regime studied rather than as a general statement about deep architectures for marketing analytics. Second, the dataset attributes are described only at a category level ("g1_", "g2_", "c_") and exact feature semantics are not disclosed by the data provider for confidentiality reasons, which limits the depth of feature-level interpretation and constrains reproducibility at the level of individual attribute analysis. Third, the evaluation is restricted to classification metrics (precision, recall, F1, accuracy, AUC); business-oriented metrics such as expected ROI uplift, conversion rate gain, customer lifetime value impact, and campaign cost per acquired customer would more directly reflect the practical utility of the predictions, but require labelled campaign-outcome data and a deployment context that are outside the scope of this study. Fourth, although the framework was designed with interpretability in mind, post-hoc explainability techniques such as SHAP value analysis or permutation feature importance were not applied to the top-performing models; integrating such analyses would strengthen the link between model predictions and actionable marketing insights. Fifth, no external dataset or temporal validation was performed: the generalization claims pertain to the held-out split of this single Kaggle source, and replication on additional retail datasets with disclosed feature semantics would be needed to establish broader generalizability. These limitations are addressed in our future work plan in Section VII.

VII. Conclusion

This study explored the application of machine learning and deep learning techniques for predicting profitable customer segments to optimize targeted marketing strategies. A comprehensive framework was developed, incorporating conventional classifiers (AdaBoost, Gradient Boosting, XGBoost, LDA, Random Forest, Naïve Bayes, SVM, and KNN) and deep learning architectures (1D ResNet and 1D DenseNet).

The experimental results demonstrated that traditional machine learning models, particularly XGBoost and Gradient Boosting, consistently outperformed deep learning models across multiple metrics. Both achieved the highest weighted F1 score of 0.62 and the highest AUC values (0.75 and 0.74, respectively). The performance advantage of traditional models is attributed to the relatively limited dataset size and class imbalance, conditions under which deep learning models struggle to generalize effectively. Among the deep learning models, DenseNet-121 outperformed all ResNet variants with a test F1 of 0.54, while deeper ResNet architectures did not improve over the simpler ResNet-18, reinforcing that model complexity should be proportional to dataset size.

XGBoost offered additional advantages in terms of execution speed and computational efficiency while maintaining comparable predictive performance to Gradient Boosting. With only 0.04M parameters and 0.15 GFLOPs for training, XGBoost required nearly six orders of magnitude fewer computations than DenseNet-121 (121232 GFLOPs). Both models demonstrated effective generalization and robust handling of imbalanced data. As traditional machine learning models, they are also more interpretable than deep learning approaches, which is a key benefit for understanding model decisions and driving actionable business insights.

The additional experiments with One-vs-Rest classification and autoencoders did not yield significant improvements over the direct multi-class approach, suggesting that the underlying class overlap in the dataset poses a fundamental challenge that cannot be easily addressed through alternative classification strategies or unsupervised feature extraction.

Future work will pursue four directions in response to the limitations identified above. First, the framework will be evaluated on larger and more diverse customer datasets, particularly those exceeding tens of thousands of instances, to characterize the data-volume regime in which deep architectures begin to outperform tree-based ensembles for this task. Second, post-hoc explainability techniques, specifically SHAP value analysis and permutation-based feature importance, will be applied to XGBoost and Gradient Boosting to produce interpretable, per-prediction attributions that marketing teams can act on, addressing the interpretability dimension that the present study only partially covers. Third, the evaluation will be extended from purely classification-oriented metrics to business-oriented metrics, including expected ROI uplift, conversion rate gain, customer lifetime value impact, and cost per acquired customer, by partnering with marketing teams to

obtain labelled campaign-outcome data that link model predictions to realized financial returns. Fourth, hybrid ensemble methods that combine the strengths of gradient boosting with neural feature extractors will be investigated as a potential route to overcoming the F1 $\approx$ 0.62 ceiling observed on the current feature set, alongside richer input features (such as marketing channel, prior campaign exposure, and product-category preference) that may help disambiguate the residual confusion involving class 0. Robust outlier detection methods will also be incorporated as part of the preprocessing pipeline to further enhance data quality.

ACKNOWLEDGMENT

First and foremost, we thank Allah Almighty, the Most Gracious, the Most Merciful, for the countless blessings that led to the successful completion of this research. We are deeply grateful to our supervisor, Prof. Ouiem Bchir, for her continuous support, patience, and dedication throughout every stage of this work. Finally, we express our deepest gratitude to our families and friends for their unwavering support, love, and encouragement.

REFERENCES

[1] N. E. Koufi, A. Belangour, and M. Sadiq, "Toward a decision-making system based on artificial intelligence for precision marketing: A case study of Morocco," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 1, p. 100250, 2024.

[2] M. Hafez, "Pioneering perspectives: Strategies and considerations in market segmentation and targeting," *SSRN Electronic Journal*, 2024.

[3] V. Kumar, A. R. Ashraf, and W. Nadeem, "AI-powered marketing: What, where, and how?" *International Journal of Information Management*, vol. 77, p. 102783, 2024.

[4] S. Adhikari, "Machine learning and customer behavior insights: Exploring the depth of predictive analytics in enhancing consumer interaction and engagement," *Journal of Empirical Social Science Studies*, vol. 8, no. 2, pp. 51–62, 2024.

[5] A. R. Thomas, "The end of mass marketing: or, why all successful marketing is now direct marketing," *Direct Marketing: An International Journal*, vol. 1, no. 1, pp. 6–16, 2007.

[6] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] B. Steele, J. Chandler, and S. Reddy, *Algorithms for Data Science*. Switzerland: Springer, 2016.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[13] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[16] A. Xu, Y. Li, and P. K. Donta, "Marketing decision model and consumer behavior prediction with deep learning," *Journal of Organizational and End User Computing*, vol. 36, no. 1, pp. 1–25, 2024.

[17] A. A. Keji, O. Fakeye, N. N. Onochie, and O. Sangotoki, "Predicting long-term deposit customers using convolutional neural network and data conversion technique," *African Scientific Reports*, p. 192, 2024.

[18] N. Nabi *et al.*, "Unleashing deep learning: Transforming e-commerce profit prediction with CNNs," *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 126–131, 2024.

[19] E. Deniz and S. Ç. Bülbül, "Predicting customer purchase behavior using machine learning models," *Information Technology in Economics and Business*, 2024.

[20] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, Inc., 2022.

[21] R. D. Krishna, M. Mahadev, S. Hariprasad, S. Abhishek, and T. Anjali, "Cultivating customer purchase intent: Leveraging machine learning for precise predictions," in *2023 12th International Conference on System Modeling and Advancement in Research Trends (SMART)*, 2023, pp. 374–379.

[22] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60 134–60 149, 2019.

[23] C. Daqing, K., and L. B. Guo, "Predicting customer profitability dynamically over time: An experimental comparative study," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer International Publishing, 2019, pp. 174–183.

[24] S. E. Saeed, M. Hammad, and A. Alqaddoumi, "Predicting customer's subscription response to bank telemarketing campaign based on machine learning algorithms," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 1474–1478.

[25] Y. Sun, D. Cheng, S. Bandyopadhyay, and W. Xue, "Profitable retail customer identification based on a combined prediction strategy of customer lifetime value," *Midwest Social Sciences Journal*, vol. 24, no. 1, pp. 104–127, 2021.

[26] Y. Sun, H. Liu, and Y. Gao, "Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model," *Heliyon*, vol. 9, no. 2, 2023.

[27] X. Xiahou and Y. Harada, "Customer churn prediction using AdaBoost classifier and BP neural network techniques in the e-commerce industry," *American Journal of Industrial and Business Management*, vol. 12, no. 3, pp. 277–293, 2022.

[28] A. Mishra and C. Sharma, "Credit card fraud detection using AdaBoost and bagging ensemble techniques," in *Proceedings of the International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, aUTHORS: please verify this entry matches the exact paper you intended to cite; the original .tex used [28] without a corresponding .bib entry. Adjust title/venue/year if needed.

[29] T.-T. Wong, N.-Y. Yang, and G.-H. Chen, "Hybrid classification algorithms based on instance filtering," *Information Sciences*, vol. 520, pp. 445–455, 2020.

[30] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 281, 2019.

[31] A. B. Musa, "Comparative study on classification performance between support vector machine and logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 4, no. 1, pp. 13–24, 2013.

[32] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.

[33] S. Moro, P. Rita, and P. Cortez, "Bank marketing – UCI machine learning repository," 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/222/bank+marketing

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[35] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decision Support Systems*, vol. 149, 2021.

[36] N. Singh, P. Singh, K. K. Singh, and A. Singh, "Machine learning based classification and segmentation techniques for CRM: a customer analytics," *International Journal of Business Forecasting and Marketing Intelligence*, vol. 6, no. 2, p. 99, 2020.

[37] S. S. Haykin, *Neural Networks and Learning Machines*. Prentice Hall/Pearson, 2009.

[38] L. Dong, C. Wang, G. Yang, Z. Huang, Z. Zhang, and C. Li, "An improved ResNet-1d with channel attention for tool wear monitor in smart manufacturing," *Sensors*, vol. 23, no. 3, 2023.

[39] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proceedings of the International Joint Conference on Neural Networks*, 2017, pp. 1578–1585.

[40] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with ResNets," Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[41] Tsiaras, Christos (contributor) and Kaggle, "Predicting profitable customer segments," Kaggle Datasets, 2024, tabular dataset, 6621 instances with grouped customer attribute features (g1_, g2_, c_) and a three-class profitability target. Accessed: 2026-05-15. [Online]. Available: https://www.kaggle.com/datasets/tsiaras/predicting-profitable-customer-segments

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*. Savannah, GA, USA: USENIX Association, 2016, pp. 265–283, software available from https://www.tensorflow.org. Version 2.x used in this work.

[43] F. Chollet *et al.*, "Keras: Deep learning for humans," Software library, 2015–2024, version 2.x used in this work. Available from https://keras.io/. [Online]. Available: https://keras.io/

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[45] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[46] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[47] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statistics and Its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[48] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[49] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

A Human-Centered Evaluation of AI-Generated Guidance: Integrated Statistical and Machine Learning Analysis with a Risk Framework for High-Stakes Domains

Omar Al-Turki[1], Felwah Alqahtani[2], Eman Alqahtani[3],
Sarah Alswedani[4], Sami Alshmrany[5], Rashid Mehmood[6]*
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia[1,5,6]
College of Computer Science, King Khalid University, Abha 62521, Saudi Arabia[2,3]
School of Computing and Information, University of Hail, Hail 55473, Saudi Arabia[4]

***Abstract*—The increasing use of large language models (LLMs) in domains requiring interpretation and judgment has raised critical questions about trust, reliability, and accountability, particularly in contexts where decisions carry significant consequences. While prior work has focused primarily on improving system performance, limited attention has been given to how users evaluate and interact with AI-generated guidance in real-world, high-stakes settings. This paper addresses this gap through a large-scale empirical investigation of public perceptions of AI-generated religious guidance in Saudi Arabia. The analysis is based on survey data collected from 572 participants and combines quantitative statistical methods with a machine learning-based pipeline for analyzing open-ended responses. The quantitative component examines patterns in trust, perceived risk, privacy concerns, credibility, and user practices, while the qualitative component employs embedding-based clustering using Bidirectional Encoder Representations from Transformers (BERT), Uniform Manifold Approximation and Projection (UMAP), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), followed by expert interpretation to derive structured parameters. The results indicate a cautious and conditional engagement with AI systems, characterized by moderate usage, low levels of trust, and strong concerns regarding reliability and source credibility. Users frequently verify AI-generated outputs and demonstrate a preference for human expert validation, particularly in complex or sensitive cases. Building on these insights, the study introduces a layered taxonomy of perceived risks spanning epistemic, reasoning, interactional, and institutional dimensions, providing a structured analytical framework for understanding how technical limitations translate into broader behavioural and governance challenges. These results highlight the importance of aligning AI system design with user expectations, emphasizing transparency, verifiability, and human oversight. The proposed taxonomy and analytical framework provide a foundation for future research and contribute to the development of governance approaches for AI systems deployed in high-stakes interpretive domains.**

***Keywords*—Large language models; AI-generated guidance; user perception; trust in AI; perceived risk; source credibility; human-AI interaction; high-stakes domains; risk taxonomy; risk framework**

*Corresponding author

I. Introduction

Artificial intelligence (AI), particularly large language models (LLMs), is increasingly embedded in domains that require not only information retrieval but also judgment, interpretation, and decision support. These systems are now consulted for guidance in areas such as healthcare, legal reasoning, education, and personal decision-making, where the consequences of errors may be significant [1], [2], [3]. As a result, questions of trust, reliability, transparency, privacy, and accountability have moved to the forefront of AI research and governance [4], [5]. Prior work has shown that LLMs can generate fluent yet inaccurate outputs, exhibit hallucinations, and influence user judgments in ways that extend beyond neutral information provision [6], [4]. In high-stakes contexts, particularly in what can be described as interpretive domains, that is, settings where responses depend on reasoning, contextual judgment, and the application of domain-specific knowledge rather than direct retrieval of factual information, these characteristics raise critical concerns regarding how users interpret AI-generated responses, how much they rely on them, and how such systems interact with existing structures of expertise and decision-making.

In response to these challenges (see e.g., [7]), emerging AI governance frameworks, such as the EU AI Act, increasingly adopt risk-based approaches that classify applications according to their potential impact on individuals and society, particularly in domains involving decision-making, access to services, and fundamental human rights [8]. Such frameworks emphasize requirements for transparency, accountability, human oversight, and reliability, especially for high-risk AI systems whose outputs may directly or indirectly influence human decisions and behaviour [4], [5], [9]. Within this context, understanding how users perceive, interpret, and evaluate AI-generated outputs becomes not only a technical concern but also a regulatory and societal imperative.

A central challenge in this landscape lies in the tension between the apparent fluency and accessibility of AI-generated responses and the underlying limitations of these systems. While LLMs can produce coherent and persuasive outputs,

concerns persist regarding their tendency to generate unsupported or misleading information, fabricate references, align responses with user expectations rather than evidence, and operate without clear accountability structures [6], [4]. These issues are amplified in domains where knowledge is not purely factual but interpretive, context-dependent, and institutionally grounded. In such settings, the role of expertise, the legitimacy of sources, and the processes through which knowledge is validated become critical factors shaping both the use and acceptance of AI systems.

Within this broader landscape, certain domains present additional challenges due to their reliance on interpretive reasoning, contextual judgment, and established authority structures. Religious inquiry represents one such domain, where responses are not purely informational but are derived through structured processes that combine source-based evidence, domain expertise, and case-specific interpretation. In the Islamic context, religious rulings (Fatwa, Fatawa) are traditionally produced by qualified scholars through systematic reasoning grounded in recognized sources and interpretive frameworks [10]. As AI systems become increasingly accessible, they are now being used by individuals to obtain guidance on religious questions, effectively introducing LLMs into a domain characterized by high requirements for accuracy, contextual sensitivity, and legitimacy of authority. This shift raises important questions about how users evaluate AI-generated responses in terms of trust, credibility, privacy, and appropriateness, and how these systems are positioned relative to established human-centred processes of knowledge validation.

Despite growing research on AI systems and their applications, existing work remains primarily focused on improving system performance or examining isolated aspects of user interaction. Studies in domains such as healthcare and legal advisory have highlighted issues related to trust, over-reliance, and the difficulty of distinguishing between expert and AI-generated responses [1], [2], [3], while recent work in the context of religious knowledge has pointed to concerns regarding correctness, governance, and alignment with scholarly standards [11], [12], [13]. However, these efforts have largely addressed technical reliability or domain-specific challenges in isolation, with limited attention to how multiple dimensions of user perception, such as trust, perceived risk, privacy, credibility, and authority, interact in shaping real-world use and acceptance of AI systems. This gap is particularly significant in high-stakes interpretive domains, where user perceptions directly influence decision-making, compliance, and the broader societal impact of AI deployment.

Across these strands of research, a more fundamental gap emerges. While prior work has examined system performance and, in some cases, user interaction in specific domains, it does not yet provide a comprehensive understanding of how AI-generated guidance is evaluated and used in high-stakes interpretive contexts. Existing efforts remain primarily oriented toward improving model outputs [14], [15], with comparatively less attention to how users perceive, interpret, and respond to these outputs in real-world settings [16], [1], [2], [3]. At the same time, emerging work in religious knowledge highlights concerns related to correctness, governance, and alignment with scholarly standards [11], [12], [13], but these aspects are often considered independently rather than as interconnected dimensions of user evaluation and system use.

In addition, there is a lack of structured conceptual frameworks that organize the diverse risks associated with AI-generated outputs into coherent and interpretable forms. Most existing approaches address issues such as hallucination, bias, or misuse as separate challenges [6], without considering how these risks propagate across technical, behavioural, and institutional levels [4]. This limitation is particularly critical in high-stakes interpretive domains, where the acceptance and impact of AI systems depend not only on technical performance but also on how users interpret outputs, assess their authority, and align them with established knowledge and decision-making processes.

To address the identified gaps, this study presents one of the first large-scale empirical investigations of public perceptions of AI-generated guidance in a high-stakes interpretive domain, using the context of religious inquiries in Saudi Arabia. The research is based on a combined quantitative and qualitative analysis of survey data collected from 572 participants, providing insight into how individuals engage with AI systems and evaluate their outputs in real-world settings.

The analysis examines user perceptions across multiple interrelated dimensions, including trust, perceived risk, privacy, credibility, and user practices, and further explores how these perceptions vary with demographic and domain-specific factors, particularly age and religious knowledge. The quantitative component employs statistical analysis of structured survey responses, including descriptive statistics, correlation analysis, and group-based comparisons using appropriate significance testing, to identify patterns in user engagement and evaluation. In parallel, the qualitative component applies a machine learning-based pipeline to open-ended responses, leveraging Bidirectional Encoder Representations from Transformers (BERT) for semantic embedding [17], followed by dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) [18] and clustering via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [19]. The resulting clusters are then interpreted through a human–AI collaborative process to derive structured parameters. This integrated statistical and machine learning-based approach enables the identification of underlying concerns and expectations while preserving both analytical rigor and contextual interpretability.

Building on these results, the study introduces a layered taxonomy of perceived risks structured across four interrelated dimensions: epistemic and knowledge integrity, contextual and reasoning limitations, human–AI interaction dynamics, and socio-technical and institutional impacts. This taxonomy provides a structured representation of how technical limitations in AI-generated outputs relate to behavioural patterns and broader governance considerations, offering a coherent framework for understanding user perception in high-stakes contexts.

This study advances a shift from system-centric evaluation toward a more comprehensive understanding of AI in practice, emphasizing the role of user perception, contextual factors, and governance requirements. By providing empirical evidence, a combined statistical and machine learning-based analytical methodology, and a structured conceptual framework, the study offers insights relevant not only to religious inquiry but also to

other high-stakes interpretive domains where trust, authority, and accountability are central to the responsible deployment of AI systems. Beyond its empirical contribution, the proposed taxonomy offers a foundation for future research by providing a systematic way to analyse and compare risks across domains, supporting the development of evaluation benchmarks, governance frameworks, and domain-specific AI systems. In this sense, the framework not only captures current user perceptions but also enables further investigation into how AI-related risks evolve across different contexts and applications.

The remainder of this study is organized as follows: Section II reviews related work and outlines the research problem. Section III describes the methodology, including the survey design and the integrated statistical and machine learning-based analysis pipeline. Section IV presents the quantitative analysis of structured responses, while Section V provides the qualitative analysis and introduces the proposed taxonomy of perceived risks and public expectations. Section VI discusses the results in relation to existing literature and AI governance considerations. Finally, Section VII concludes the study and outlines directions for future research.

II. Literature Background

Recent research has increasingly explored the intersection of artificial intelligence and Islamic knowledge systems, particularly in the context of religious guidance generation and question answering (QA). Early efforts focused on building structured datasets and resources to support AI-driven religious applications. For instance, Alyemny et al. [14] introduced one of the first Arabic religious rulings datasets, compiled from authoritative sources such as the "General Presidency of Scholarly Research and Ifta" in Saudi Arabia [20], prominent scholars including Ibn Othaimin [21], and online repositories like FatwaPedia [22]. This work laid the foundation for developing data-driven AI systems in the Islamic domain.

Building on such resources, several studies have proposed AI-based systems to enhance the reliability of religious question answering. Alan et al. [15] developed MufassirQAS, a retrieval-augmented generation (RAG) system that integrates the Qur'an, Sunnah, and core Islamic teachings as grounding sources. Their findings demonstrated improved response clarity and reduced ambiguity compared to general-purpose models such as ChatGPT. Similarly, Mohammed et al. [16] proposed an advanced RAG-based framework incorporating a Flash re-ranker to reduce hallucinations in AI-generated religious rulings (Fatawa). Their evaluation across multiple models, including AceGPT [23], Gemini-1.5 [24], and SILMA [25], showed that combining RAG with re-ranking significantly improved factual accuracy, albeit at the cost of increased latency.

In parallel, recent work has examined critical reliability aspects of AI systems in religious contexts. Abstention, the ability of a model to refrain from answering when uncertain, has been identified as a crucial requirement in sensitive domains [26]. Atif et al. [27] introduced FiqhQA, a benchmark designed to evaluate LLM performance across different Islamic schools of thought (Madhhabs), highlighting the importance of doctrinal diversity and controlled abstention in AI-generated religious responses. These studies are collectively summarized in Table I.

TABLE I. Summary of AI-based Religious QA Studies

Study	Summary	Research Gap
MufassirQAS [15]	RAG-based QA system using Qur'an, Hadith, and Islamic texts to improve accuracy, reduce hallucinations, and provide source transparency.	Limited generalization, bias issues, and lack of multilingual and contextual depth.
Aftina [16]	RAG framework with re-ranking to improve retrieval quality and reduce hallucinations in AI-generated religious rulings.	Depends on dataset quality; challenges in Arabic understanding and limited domain data.
FiqhQA Benchmark [27]	Introduces a benchmark dataset covering Islamic rulings across four schools of thought and evaluates LLM accuracy and abstention behavior. Shows performance varies by language and model, with weaker results in Arabic.	Lacks standardized evaluation methods, struggles with multilingual reasoning, and highlights need for better abstention, transparency, and alignment with jurisprudential diversity.

While these studies focus primarily on improving the technical performance and reliability of AI systems, a growing body of research has begun to examine how users perceive and interact with AI-generated advice in high-stakes domains. In the medical field, Shekar et al. [1] found that non-expert users often struggle to distinguish between AI-generated and expert-generated responses, and may exhibit high willingness to follow inaccurate AI advice. In the legal domain, Schneiders et al. [2] reported that participants frequently preferred AI-generated advice over professional legal guidance when the source was not disclosed. Similarly, Seabrooke et al. [3] showed that while prior usage of AI for legal advice remains limited, a substantial proportion of users expressed willingness to adopt such tools in the future, particularly for less sensitive topics. Taken together, these studies are summarized in Table II.

TABLE II. Summary of User Perception Studies in High-Stakes AI Domains.

Study	Summary	Research Gap
Medical Domain [1]	Users cannot distinguish AI from doctors and often trust AI responses, even when inaccurate, with willingness to follow harmful advice.	Overtrust in AI and poor error detection; lacks safeguards to prevent harmful reliance.
Legal Domain [2]	Users show higher willingness to act on AI-generated legal advice, especially when the source is unknown, despite being able to distinguish it.	Preference bias toward AI advice without mechanisms to ensure safe decision-making.
Legal Domain [3]	Low current use of AI for legal advice, but strong future willingness, varying by legal domain.	Lacks evaluation of real-world risks, decision quality, and consequences of AI reliance.

Despite these advancements, limited research has investigated public perceptions of AI tools within the religious domain, especially in Islamic contexts where authority, trust, and source credibility play a central role. Existing studies largely emphasize system design, dataset construction, or technical evaluation, with insufficient attention to how users assess trust, risk, and legitimacy when interacting with AI-generated religious guidance.

More recently, emerging studies have begun to explore AI within the context of religious knowledge, user perception, and ethical governance. Alam et al. [11] conducted a mixed-methods investigation of user and expert evaluations of AI-generated religious content, revealing significant challenges in users' ability to assess correctness and reliability. In the Islamic domain, Priantina et al. [12] examined the integration of AI into formulation of religious rulings, highlighting both efficiency gains and critical concerns related to automating

sacred decision-making and weakening scholarly reasoning. Furthermore, a recent large-scale survey by Bhatia et al. [13] synthesized over 160 studies on AI for Islamic knowledge, identifying key gaps in trustworthiness, evaluation frameworks, and governance mechanisms. These findings collectively emphasize that, despite rapid technological advancements, the human-centered dimensions of trust, risk perception, and authority remain underexplored. A summary of these studies is presented in Table III.

TABLE III. SUMMARY OF AI IN RELIGIOUS AND ISLAMIC KNOWLEDGE STUDIES

Study	Summary	Research Gap
[11]	Users prefer AI-generated religious answers (81%) despite claiming trust in scholars; experts identify major quality issues.	User–expert misalignment and limited ability to detect errors; need expert-guided evaluation.
[13]	Comprehensive review of AI systems for Islamic knowledge, highlighting growth of RAG, evaluation methods, and alignment challenges.	Fragmented research, lack of unified evaluation standards, and limited handling of pluralism and trust.
[12]	AI improves efficiency and scalability of formulation of religious rulings but raises concerns about replacing scholarly reasoning.	Ethical risks, lack of governance frameworks, and need for human oversight in decision-making.

A. Research Problem

While the preceding discussion highlights substantial progress in developing AI systems for religious knowledge, including advances in dataset construction, retrieval-augmented generation, and benchmark evaluation, it also reveals important limitations in how these systems are assessed beyond technical performance. In particular, existing research has largely focused on improving accuracy and reducing hallucination in AI-generated outputs [14], [15], [16], [27], with comparatively limited attention to how such outputs are interpreted and evaluated by users in real-world settings.

In parallel, a growing body of work has examined user interaction with AI systems in high-stakes domains such as healthcare and legal advisory, highlighting challenges related to trust, over-reliance, and the difficulty of distinguishing between expert and AI-generated responses [1], [2], [3]. More recent research in the context of religious knowledge has identified additional concerns, including discrepancies between user perception and expert evaluation, risks associated with automating interpretive reasoning, and the absence of comprehensive governance and evaluation frameworks [11], [12], [13]. Despite these advances, existing research remains siloed, with studies typically addressing either system performance or isolated aspects of user perception, rather than providing a comprehensive understanding of how multiple factors interact in shaping user evaluation and use of AI-generated guidance.

A key limitation in the current literature is the lack of an integrated perspective that connects core dimensions of user perception, including trust, perceived risk, privacy, credibility, and user practices, with broader considerations of authority and governance. In domains characterized by interpretive reasoning and contextual judgment, such as religious inquiry, these dimensions are inherently interdependent. However, prior work has largely examined them in isolation, without capturing how concerns related to accuracy, contextual understanding, interaction behaviour, and institutional authority collectively influence the acceptance and use of AI systems.

Furthermore, there is a lack of structured conceptual frameworks that organize the diverse risks associated with AI-generated outputs into coherent and interpretable categories. Existing approaches typically address issues such as hallucination, bias, or misuse as separate challenges, without considering how these risks propagate across technical, behavioural, and institutional levels. This limits the ability to develop a holistic understanding of AI-related risks in high-stakes interpretive contexts.

To the best of our knowledge, large-scale empirical evidence on how users evaluate AI-generated guidance in high-stakes interpretive domains remains limited globally. This gap is particularly pronounced in context-specific environments such as Saudi Arabia, where domain expertise, cultural expectations, and established authority structures play a central role in shaping user behaviour.

To address these gaps, this study presents one of the first large-scale empirical investigations of public perceptions of AI-generated religious guidance in Saudi Arabia, based on a combined quantitative and qualitative analysis of survey data. The work examines how users engage with AI systems and how they evaluate outputs across multiple dimensions, including trust, perceived risk, privacy concerns, and source credibility. In addition, it employs an integrated statistical and machine learning-based analysis to derive structured parameters from open-ended responses, enabling the identification of underlying concerns and expectations through clustering and expert interpretation. Building on this, the study introduces a layered taxonomy of perceived risks structured across four interrelated dimensions: epistemic and knowledge integrity, contextual and reasoning limitations, human–AI interaction dynamics, and socio-technical and institutional impacts. This structured representation captures how concerns originate at the level of AI-generated outputs and propagate through reasoning, user interaction, and institutional contexts, providing a coherent framework for understanding the interdependencies between technical limitations, user behaviour, and governance considerations.

To operationalize these objectives, the study addresses the following research questions:

RQ1. How do individuals in Saudi Arabia engage with and understand AI tools in the context of religious inquiries?

RQ2. How are trust, perceived risk, privacy, and credibility evaluated when AI systems are used to generate religious guidance?

RQ3. How do demographic and domain-specific factors, particularly age and religious knowledge, influence these perceptions and related user practices?

III. METHODOLOGY AND DESIGN

We conducted a survey to examine participants' perceptions regarding the use of AI for generating religious content. The survey was designed to capture both demographic characteristics and participants' perceptions toward AI-generated religious guidance.

The questionnaire consisted of two main parts: 1) Demographic information was collected from participants, including age, gender, level of education, religious background, AI usage, and whether they usually consult scholars or official Islamic websites when seeking religious information. 2) Participants' views on AI-generated religious content. We grouped the related questions, as shown in Table IV. This section captured participants' trust in AI-generated religious content, perceived risks of relying on AI for religious guidance, and perceptions regarding the credibility of AI-generated sources. Open-ended questions were included to allow participants to elaborate on their concerns and provide suggestions regarding the appropriate role of AI in religious contexts.

Note that the questionnaire items were developed based on themes identified in prior literature on trust, credibility, risk perception, and AI usage, and were reviewed for clarity and contextual appropriateness prior to distribution. Internal consistency of the quantitative constructs was subsequently assessed using Cronbach's alpha.

TABLE IV. QUESTIONNAIRE USED IN THE RESEARCH

Variables	Questions	Question type
Familiarity with AI	How often do you request a religious ruling from AI?	Frequency Scale (1–5)
	How deep is your understanding of AI tools?	Ranked categories from high to low understanding
Trust in AI	I trust AI-generated religious rulings	Agreement scale (1–5)
	I believe AI-generated religious rulings are accurate	Agreement scale (1–5)
Perceived Privacy	I prefer to seek religious rulings from AI about things I might feel hesitant to ask someone	Agreement scale (1–5)
Perceived Risk	I feel concerned about the implications of using AI for religious rulings	Agreement scale (1–5)
	I believe that asking AI for religious rulings weakens the relationship between society and scholars	Agreement scale (1–5)
	I believe that seeking religious rulings from AI may reduce the level of concern for the accuracy of religious rulings	Agreement scale (1–5)
Perceived Credibility	I prefer that the AI-generated religious rulings be linked to well-known sources (e.g., the Council of Senior Scholars)	Agreement scale (1–5)
	To what extent do you personally verify the reliability of the source provided by AI before relying on its answer?	Frequency Scale (1–5)
Behavioral Preference	If you encounter a religious ruling from AI that contradicts the religious ruling of a scholar you trust, which one would you follow?	Choice Question
Use Intention	Would you consider asking AI for a religious ruling in the future?	Yes and No question

A. Data Collection

We collect data for our research by conducting a large-scale online study. We recruited participants using different approaches such as email (from both academic and non-academic environments) and from social networks. At the beginning of the research, 573 participants read and gave their consent to participate in the research. The inclusion criteria require participants to be 18 years or older. We included a total of 572 responses in this analysis after we filtered out incomplete responses.

B. Participants

A total of 572 participants (age range of most (74.6%) of our participants was 18-44 years; 169 males and 326 females) who had used AI before. Most of our participants (94.4%), turn to scholars to obtain religious rulings. As shown in Table V, the participants were diverse in terms of age, gender, education level, and religious (Fiqh) background.

TABLE V. DEMOGRAPHIC CHARACTERISTICS OF PARTICIPANTS

Age	18–28 (211, 36.9%), 29–44 (192, 33.6%), 45–60 (152, 26.6%), above 61 (17, 3%)
Gender	Male (188, 32.9%), Female (384, 67.1%)
Educational level	High school (88, 15.4%), Undergraduate (341, 59.6%), Graduate (132, 23.1%), Other (11, 1.9%)
Religious Background	Basic (167, 29.2%), Intermediate (256, 44.8%), Advanced (149, 26%)
Seek Religious Rulings	Scholars (540, 94.4%), Official Islamic websites (32, 5.6%)

C. Data Analysis

Fig. 1 presents the overall architecture adopted in this research, illustrating how quantitative statistical analysis and qualitative BERT-based modeling are integrated to generate a comprehensive understanding of public perceptions of AI-generated religious rulings. The framework begins with survey-based data collection, capturing both structured (closed-ended) and unstructured (open-ended) responses from participants. These two data streams are then processed through parallel analytical pipelines.

The quantitative pipeline focuses on structured responses and applies statistical techniques to measure key constructs such as trust, perceived privacy and risk, and credibility, as well as to examine relationships and demographic effects. In parallel, the qualitative pipeline processes open-ended responses using a BERT-based modeling approach to extract latent themes, enabling the identification of underlying concerns, expectations, and recommendations.

The outputs of both pipelines are subsequently integrated through a triangulation process, ensuring complementarity and cross-validation between statistical patterns and thematic insights. This integration step forms the core of the framework, allowing the research to move beyond isolated results toward a unified interpretation of user perceptions. The final stage translates these integrated insights into a structured understanding of perception patterns, key concerns, and evidence-based implications for the design and governance of AI systems in religious contexts.

While Fig. 1 provides a high-level overview of this analytical architecture, the detailed methodological components of the quantitative and qualitative pipelines are described in the following sections.

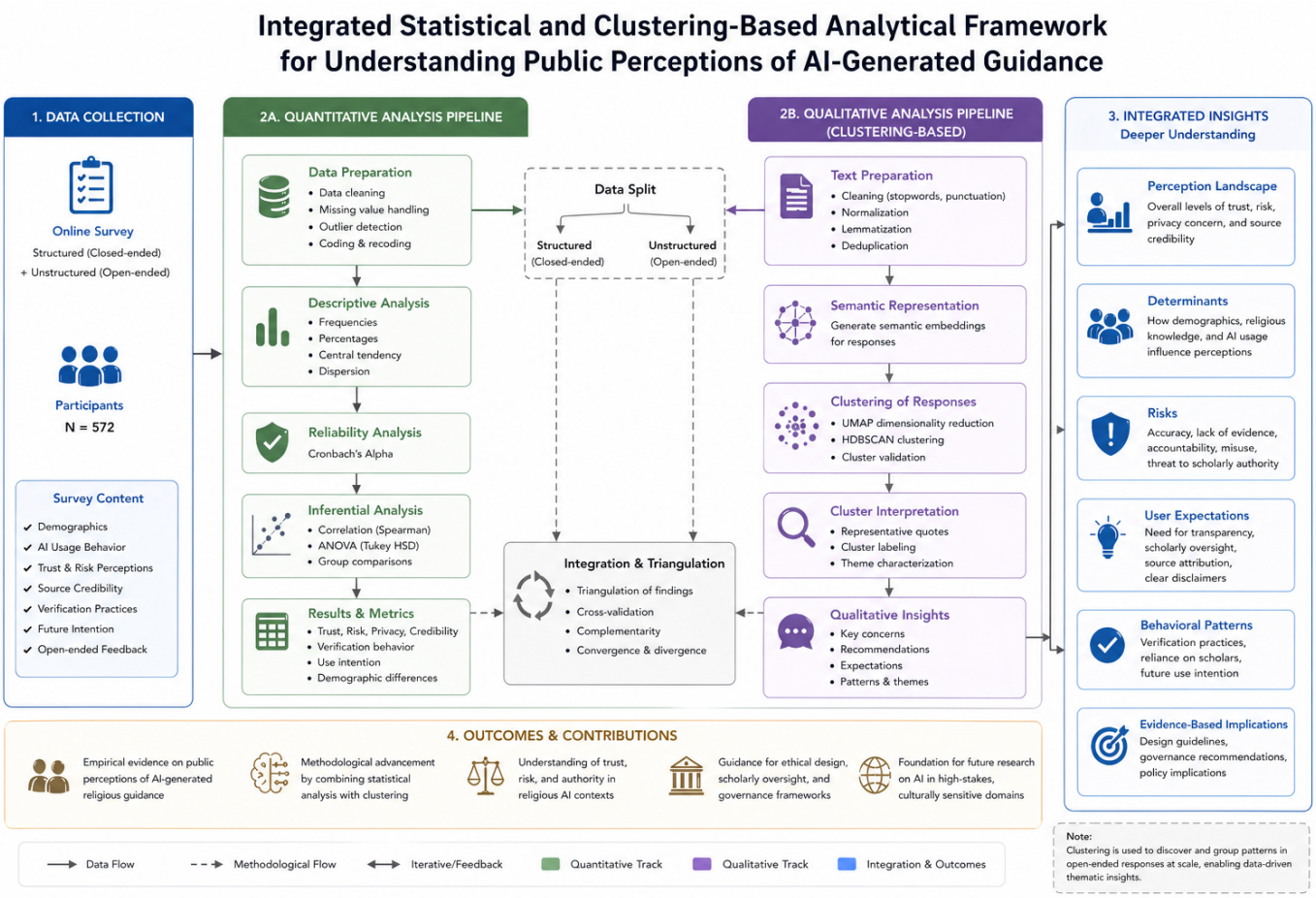


Fig. 1. Integrated statistical and clustering-based analytical framework for examining public perceptions of AI-generated guidance. The figure illustrates the end-to-end methodology combining quantitative statistical analysis of structured survey responses with embedding-based clustering of open-ended responses, followed by an integration and triangulation process. This framework enables the identification of patterns in trust, perceived risk, privacy, credibility, and user behavior, while deriving structured parameters and a layered taxonomy of risks across epistemic, reasoning, interactional, and socio-technical dimensions, supporting a unified interpretation of user perceptions in high-stakes interpretive domains.

D. Quantitative Data Analysis

To analyze the quantitative data and address the research questions, established statistical analysis techniques were employed using SPSS. Descriptive statistics were calculated to summarize the demographic characteristics of the participants. Cronbach's alpha was used to assess internal reliability when using two or three items on a 5-point Likert scale (ranging from Strongly Disagree to Strongly Agree) to measure one variable. If Cronbach's alpha $>= 0.70$, we compute the mean score of the items for the variable.

A one-way ANOVA tests were conducted to investigate whether demographic factors such as age and religious background influence the use of AI by participants when seeking religious rulings. A Tukey HSD post hoc test was conducted after the significant ANOVA result to examine differences between the groups.

E. Parameter Discovery and Qualitative Data Analysis

Open-ended survey responses were analyzed using a structured qualitative modelling pipeline based on BERTopic [17], which integrates transformer-based embeddings, dimensionality reduction, and density-based clustering to identify semantically coherent themes in textual data. Sentence-level embeddings were generated using the Arabic language model CAMeL-Lab/bert-base-arabic-camelbert-mix [28]. These embeddings were projected into a lower-dimensional space using UMAP [18] to preserve local semantic structure while enabling effective clustering. HDBSCAN [19] was then applied to automatically identify clusters without requiring a predefined number of topics.

The analysis was conducted separately on the concerns and suggestions datasets to capture distinct dimensions of user perceptions, namely perceived risks and proposed safeguards. This separation ensures that emergent themes are not conflated during the initial clustering process. Following clustering, topic representations were refined through post-processing, including stopword removal and inspection of representative responses, to improve interpretability while preserving contextual integrity. The resulting clusters were then comparatively examined across datasets to identify recurring and complementary themes.

Finally, clusters were abstracted into higher-level parame-

ters through an iterative human-guided consolidation process, ensuring semantic coherence, conceptual distinctiveness, and alignment with the overall analytical framework. This process resulted in a structured taxonomy of user-perceived risks and considerations associated with AI-generated guidance.

IV. Quantitative Analysis of Structured Survey Responses

Here, in this section, we present the results of the quantitative analysis of the survey data. We begin by reporting descriptive statistics and key patterns related to AI usage, trust, perceived risk, privacy, and credibility, followed by inferential analysis examining relationships between variables and differences across demographic and domain-specific groups. These results establish the empirical basis for the qualitative analysis presented in Section V.

A. Participants' Familiarity and Understanding of AI Tools

To understand the participants' familiarity with AI tools, we first examined how often they used AI for religious rulings, then we assessed the depth of their understanding of AI tools. As shown in Fig. 2, 31.3% of participants stated that they sometimes used AI to obtain religious content, while 30.8% reported that they never use AI for this purpose. Moreover, 18.9% participants reported rarely using AI and only 19% of participants indicated that they use AI often or always. These results show that participants use AI tools sometimes rather than regularly for religious inquiries. It suggests that there is moderate adoption of AI-generated religious rulings, but some people may hesitate or have limited trust in them.

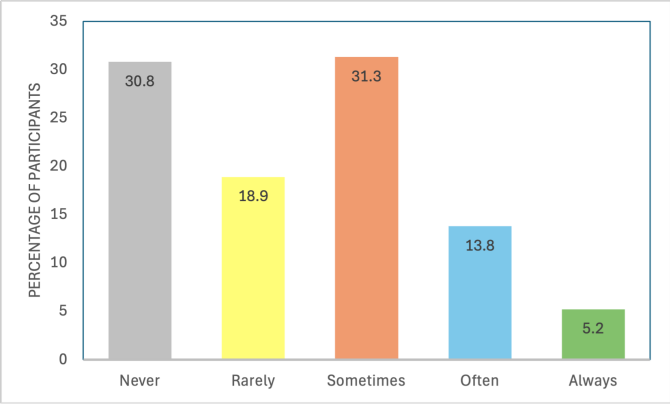


Fig. 2. Frequency of AI use among participants.

Regarding the understanding of AI tools by participants, Fig. 3 shows that 40.4% of participants stated that they have a basic understanding of AI, while 32.7% indicated having only a general knowledge of AI. Moreover, 14.9% of the participants reported that they had no knowledge of AI and only 12.1% stated that they had a deep knowledge of AI tools. The results show that most participants know about AI tools, but few fully understand how these tools work.

To understand the relationship between AI understanding and its frequency of use for religious rulings, we run a Spearman correlation analysis. The results show no significant relationship between AI understanding and its frequency of use for religious rulings (r = 0.035, p = 0.400), indicating that the level of AI understanding does not significantly impact the frequency of using AI for religious rulings. Table VI presents the results on AI familiarity, understanding, and their relationship.

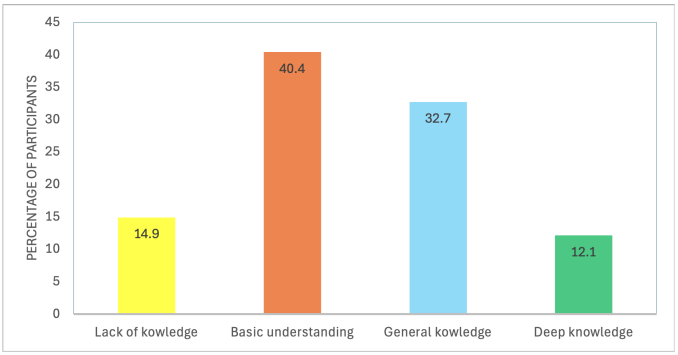


Fig. 3. Percentage of participants' understanding of how AI tools work.

TABLE VI. Summary of AI Familiarity and Understanding

Variable	Results	Interpretation
Familiarity with AI	Often or always (19%) sometimes (31.3%) rarely (18.9%) never (30.8%)	There is moderate adoption of AI-generated religious rulings.
Understanding of AI Tools	Basic understanding (40.4%) general knowledge (32.7%) no knowledge (14.9%) deep knowledge (12.1%)	Most participants are aware of AI tools, but only a few have a deep understanding of how they work.
Understanding AI & Usage	r = 0.035, p= 0.400 relationship between AI understanding and its frequency of use for religious rulings	level of AI understanding does not significantly impact the frequency of using AI for religious rulings.

B. Trust in AI

There are two items that measure the trust in AI for religious rulings, First we examined the reliability of the two items using Cronbach's alpha. The result shows an acceptable internal consistency ($\alpha = 0.785$), indicating that the two items reliably measure the trust in AI for religious rulings. Then we calculated the mean and standard deviation of trust in AI for religious rulings. The results show a mean score of 1.85 (SD = 0.96) based on 572 responses. It suggests that participants show a low level of trust in AI-generated religious rulings.

To understand whether Trust in AI for religious rulings impact the intention to use AI in the future, we run a Spearman correlation analysis. We found a moderate and statistically significant positive relationship (r = 0.499, $p < 0.001$). This indicates that participants with higher levels of trust in AI-generated religious rulings are more likely to consider using AI for religious rulings in the future and vice versa.

We also performed a Spearman correlation analysis to examine the relationship between trust in AI-generated religious rulings and the preference of participants when AI-generated religious rulings contradict those of trusted scholars. The results show a weak but statistically significant positive relationship (r = 0.151, $p < 0.001$). This shows that participants with higher levels of trust in AI-generated religious rulings

are more likely to consider following AI-generated religious rulings, even when they contradict the opinions of trusted scholars. Table VII presents the results of trust in AI-generated religious rulings and its association with use intention and behavioral preference.

TABLE VII. TRUST IN AI-GENERATED RELIGIOUS RULINGS

Variable	Results	Interpretation
Trust in AI for Religious Rulings	M = 1.85, SD = 0.96 based on 572 responses	Participants show a low level of trust in AI-generated religious rulings.
Trust in AI and Use Intention	$r = 0.499$, $p < 0.001$ between trust in AI and use intention	Participants with higher levels of trust in AI-generated religious rulings are more likely to consider using AI for religious rulings in the future, and vice versa.
Trust in AI and Behavioral Preference	$r = 0.151$, $p < 0.001$ between trust in AI and behavioral preference	Participants with higher levels of trust in AI-generated religious inquiries are more likely to consider following AI-generated religious rulings, even when they contradict the opinions of trusted scholars.

C. Perceived Privacy, Risk, and Credibility of AI for Religious Inquiries

We measure the perceived privacy of AI for religious rulings by understanding if participants prefer to seek religious rulings from AI for questions they might feel hesitant to ask others. The results show that a majority of participants do not prefer to seek religious rulings from AI for questions they might feel hesitant to ask others. Fig. 4 shows that 39% of participants strongly disagreed and 13.8% disagreed, while 23.1% remained neutral. Only 24.1% of participants expressed agreement. This indicates that most participants are hesitant to rely on AI for sensitive religious questions.

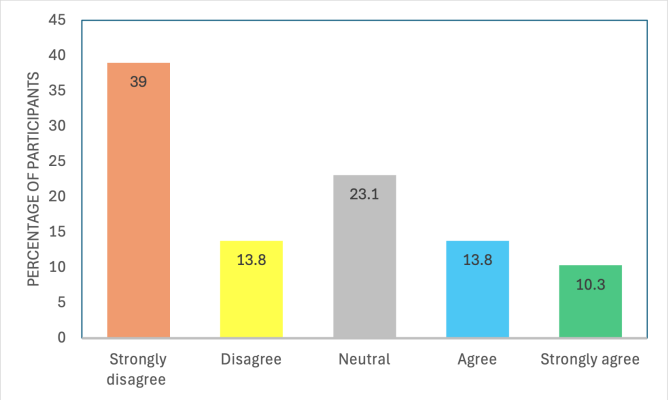


Fig. 4. Percentage of participants using AI for sensitive religious questions.

We examined the reliability of the three-item scale measuring perceived risk of AI using Cronbach's alpha. The results indicated acceptable internal consistency ($\alpha = 0.737$), suggesting that the three items reliably measure the perceived risk of AI. The mean clarifies that participants have a high level of perceived risk regarding the use of AI for religious rulings (M = 3.99, SD = 1.02, N = 572). This suggests that participants generally agree with statements expressing concerns about the implications of AI-generated religious rulings.

Perceived credibility of AI-generated religious rulings was examined using two indicators: Participants' tendency to verify the reliability of sources provided by AI and their preference for AI-generated religious rulings linked to well-known scholarly authorities. Regarding participants' tendency to verify the reliability of sources provided by AI, the results show that participants consider the importance on checking the credibility of AI-generated religious rulings. As shown in Fig. 5, they always verify the reliability of sources (28.8%), followed by 24.8% who indicated that they often verify them. Overall, 51.4% of participants reported that they often or always verify the reliability of sources provided by AI, highlighting the importance of credible references when relying on AI-generated religious rulings.

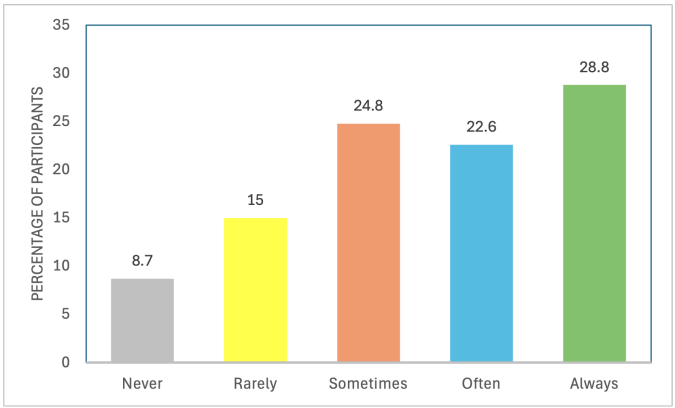


Fig. 5. Percentage of participants verifying the reliability of sources provided by AI.

Regarding participants' preference for AI-generated religious rulings linked to well-known scholarly authorities, the results show strong support for linking AI-generated religious rulings to well-known scholarly sources. Fig. 6 shows that the majority of participants strongly agreed (65.6%) to link AI-generated religious rulings to well-known scholarly sources, while 14% agreed, resulting in 79.6% expressing agreement overall. These results highlight the importance of authoritative references in enhancing the perceived credibility of AI-generated religious guidance.

To find the relationship between Participants' tendency to verify the reliability of sources provided by AI and their preference for AI-generated religious rulings linked to well-known scholarly authorities, we run a Spearman correlation analysis. The results revealed a weak but statistically significant positive relationship (r = 0.103, p = 0.014), that people who verify AI sources more frequently show a slightly stronger preference for AI-generated religious rulings supported by authoritative references. As illustrated in Table VIII, participants' perceptions of privacy, risk, and credibility shape their views on using AI for religious inquiries.

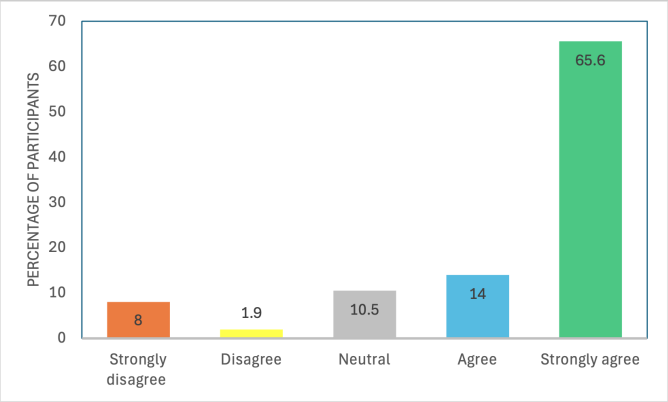


Fig. 6. Percentage of participants preferring for AI-generated religious rulings linked to well-known scholarly authorities.

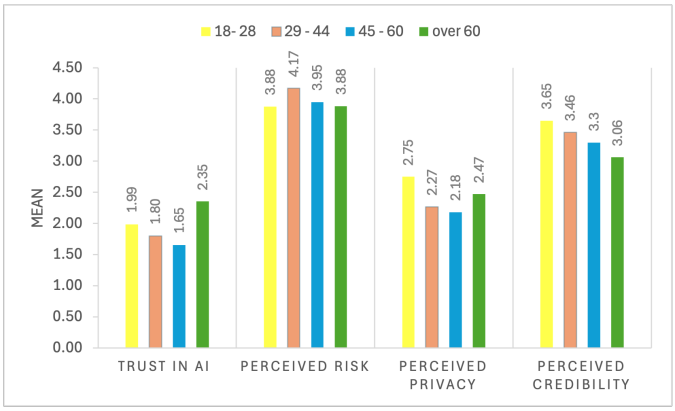


Fig. 7. Participants' perceptions based on age (Mean Values).

TABLE VIII. PERCEIVED PRIVACY, RISK, AND CREDIBILITY OF AI FOR RELIGIOUS INQUIRIES.

Variable	Results	Interpretation
Perceived Privacy	strongly disagreed (39%) disagreed (13.8%) neutral (23.1%) agree (13.8%) strongly agree (10.3%)	Most participants are hesitant to rely on AI for sensitive religious inquiries.
Perceived Risk	M = 3.99, SD = 1.02 based on 572 responses	Participants have a high level of perceived risk regarding the use of AI for religious inquiries.
Perceived Credibility	1- Tendency to verify reliability of sources: often or always (51.4%) sometimes (24.8%) rarely (15%) never (8.7%) 2- Preference for linking AI to scholarly authorities: strongly agree (65.6%) agree (14%) neutral (10.5%) disagree (1.9%) strongly disagree (8%) Relationship between (1) and (2): $r = 0.103$, $p = 0.014$	The importance of credible and authoritative references improves the perceived credibility of AI-generated religious inquiries. Participants who frequently verify sources are more likely to prefer linking AI to well-known scholarly authorities.

D. The Impact of Age Toward Using AI-Generated Religious Rulings

To examine whether age influence participants toward AI-generated religious rulings, we conducted one-way ANOVA analyses followed by a Tukey HSD post hoc test to identify differences between age groups. The overall trends are illustrated in Fig. 7.

1) Trust in AI: The results showed a statistically significant difference in trust levels [F(3,568) = 5.342, p = 0.001], indicating that trust in AI-generated religious rulings varies significantly among the age groups.The results of Tukey HSD post hoc showed that participants age range of 18-28 years reported significantly higher trust in AI-generated religious rulings compared to age group (45 - 60) (p = 0.006). Additionally, participants age over 60 showed significantly higher trust levels than participants age range 45 - 60 (p = 0.022). No other significant differences between age groups were observed.

2) Perceived risk: The results revealed a statistically significant difference in perceived risk among the age groups [F(3,568) = 3.073, p = 0.027], indicating that perceived risk of AI-generated religious rulings varies significantly between the groups. The results of Tukey HSD post hoc showed participants age range 29 - 44 perceived significantly higher risk of AI-generated religious rulings than those age range 18-28 (p = 0.019), while no other significant differences between age groups were found.

3) Perceived privacy: The results revealed a statistically significant difference among the age groups [F(3,568) = 6.481, $p < 0.001$], indicating that the willingness to consult AI for sensitive religious questions varies significantly between the groups. The results of Tukey HSD post hoc show that participants in age range (18-28) reported significantly higher willingness to consult AI for sensitive religious questions compared with participants 29-44 (p = 0.002) and participants age 45-60 ($p < 0.001$). No significant differences were observed between the remaining age group comparisons.

4) Perceived credibility: The results revealed a statistically significant difference between the age groups [F(3,568) = 2.905, p = 0.034], indicating that the tendency to verify AI-generated sources varies significantly among the groups. Post hoc comparisons using the Tukey HSD test revealed that participants in age range (18-28) verify AI-generated sources significantly more often than those in age (29-44) (p = 0.049), while no other significant differences were found among the age groups.

Overall, age has a significant influence on Participants' perception toward AI-generated religious rulings. Post hoc analysis indicated that participants in the youngest age group reported greater willingness to seek religious rulings from AI for questions they might feel hesitant to ask others, compared with some older age groups. Additionally, younger participants showed slightly higher levels of source verification behaviour than some other age groups. As shown in Table IX, age differences play a significant role in shaping trust, perceived risk, privacy concerns, and credibility of AI-generated religious information.

TABLE IX. AGE DIFFERENCES IN TRUST, RISK, PRIVACY, AND CREDIBILITY.

Variable	Results	Interpretation
Trust in AI	F = 5.342 $p = 0.001$ Post hoc (Tukey): (18–28) > (45–60): $p = 0.006$ (60+) > (45–60): $p = 0.022$	Participants aged 18–28 showed higher trust and greater willingness to use AI for sensitive religious inquiries, along with more frequent verification of AI-generated sources.
Perceived Risk	F = 3.073 $p = 0.027$ Post hoc (Tukey): 29–44 > 18–28: $p = 0.019$	Participants aged 29–44 reported higher perceived risk than younger participants.
Perceived Privacy	F = 6.481 $p = 0.001$ Post hoc (Tukey): 18–28 > 29–44: $p = 0.002$ 18–28 > 45–60: $p = 0.001$	Younger participants (18–28) showed higher concern about privacy compared to older groups.
Perceived Credibility	F = 2.905 $p = 0.034$ Post hoc (Tukey): (18–28) > (29–44): $p = 0.049$	Younger participants tend to perceive AI-generated religious information as more credible compared to middle-aged participants.

E. The Impact of Religious Background Toward Using AI-Generated Religious Content

To examine whether religious background influences participants toward AI-generated religious rulings, we conducted one-way ANOVA analysis followed by a Tukey HSD post hoc test to identify differences between groups. The overall differences across religious background levels are illustrated in Fig. 8.

1) Trust in AI: The results indicated a statistically significant difference in trust levels [F(2,569) = 9.834, p $<$ 0.001], suggesting that trust in AI-generated religious rulings varies significantly across people with different religious background. The results of a Tukey HSD post hoc text showed that participants with basic level of religious background reported significantly higher trust compared with those with intermediate (p = 0.003) and advanced levels of religious knowledge (p $<$ 0.001). No significant difference was found between intermediate and advanced religious knowledge levels.

2) Perceived risk: The results revealed a statistically significant difference in perceived risk [F(2,569) = 9.778, p $<$ 0.001], indicating that perceived risk varies significantly across people with different level of religious background. The results of Tukey HSD post hoc showed participants with basic level of religious background reported significantly lower perceived risk compared to those with intermediate (p = 0.003) and advanced religious knowledge (p $<$ 0.001). No significant difference was observed between the intermediate and advanced groups.

3) Perceived privacy: The results revealed a statistically significant difference among the participants with different level of religious background [F(2,569) = 4.113, p = 0.017], indicating that willingness to ask AI for sensitive religious questions varies significantly between groups. The results of a Tukey HSD post hoc show that participants with basic level of religious background reported significantly higher willingness to ask AI for sensitive religious questions compared with those with higher religious background (p = 0.021) and intermediate level of religious knowledge (p = 0.05). No statistically significant differences were observed between intermediate and advanced levels of religious background.

4) Perceived credibility: The results revealed a statistically significant difference between the groups [F(2,569) = 7.611, p $<$ 0.001], indicating that the tendency to verify AI-generated sources varies significantly between the groups. Post hoc comparisons using the Tukey HSD test revealed that participants in advanced religious knowledge level reported significantly higher verification of AI sources compared with those in basic religious knowledge level (p $<$ 0.001) and intermediate religious knowledge level (p = 0.036). No significant difference was found between basic religious knowledge level and intermediate religious knowledge level.

Overall, participants with basic levels of religious knowledge reported significantly higher trust in AI-generated religious rulings, greater willingness to ask AI for sensitive religious questions, and lower perceived risk compared to those with intermediate and advanced levels of religious knowledge. However, participants with advanced religious knowledge were more likely to verify the reliability of sources provided by AI, indicating greater scrutiny when evaluating AI-generated religious rulings. As illustrated in Table X, differences in religious knowledge levels significantly influence trust, perceived risk, privacy, and credibility of AI.

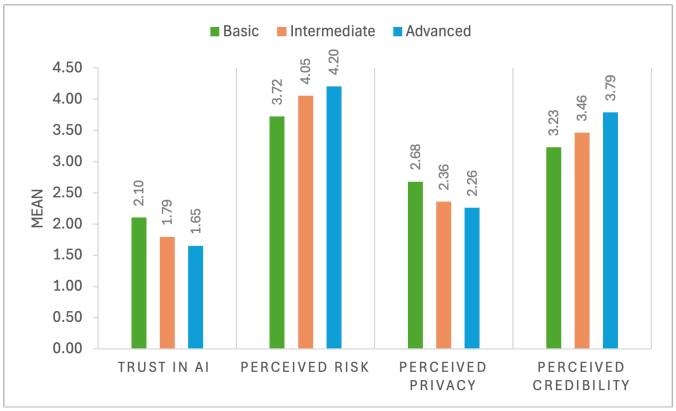


Fig. 8. Participants' perceptions based on the level of religious background (Mean Values).

V. QUALITATIVE ANALYSIS OF OPEN-ENDED SURVEY RESPONSES

The qualitative analysis of participants' open-ended responses provides deeper insight into how individuals interpret and evaluate the use of AI systems in high-stakes, interpretive contexts. While the quantitative results establish broad patterns related to trust, perceived risk, privacy perceptions, and credibility, the qualitative analysis reveals the underlying reasoning, concerns, and expectations that shape these perceptions.

Using a machine learning–based text analysis pipeline, responses were clustered and iteratively refined to extract coherent themes. These themes were interpreted as structured parameters capturing participants' perceptions. To ensure conceptual clarity and avoid overlap, the resulting parameters were organized into a hierarchical taxonomy consisting of four macro-dimensions (macro-parameters). These macro-parameters reflect distinct layers within the broader socio-technical system in which AI operates, spanning multiple

TABLE X. DIFFERENCES BASED ON RELIGIOUS KNOWLEDGE LEVELS

Variable	Results	Interpretation
Trust in AI	F = 9.834 $p = 0.001$ Post hoc (Tukey): Basic > Intermediate: $p = 0.050$ Basic > Advanced: $p = 0.021$	Participants with basic religious knowledge showed higher trust, lower perceived risk, and greater willingness to use AI for sensitive religious inquiries compared to those with intermediate and advanced levels.
Perceived Risk	F = 9.778 $p = 0.001$ Post hoc (Tukey): Basic < Intermediate: $p = 0.003$ Basic < Advanced: $p = 0.001$	Participants with advanced religious knowledge reported higher perceived risk compared to those with basic knowledge.
Perceived Privacy	F = 4.113 $p = 0.017$ Post hoc (Tukey): Basic > Intermediate: $p = 0.050$ Basic > Advanced: $p = 0.021$	Participants with lower religious knowledge expressed greater privacy concerns than those with higher knowledge levels.
Perceived Credibility	F = 7.611 $p = 0.001$ Post hoc (Tukey): Advanced > Basic: $p < 0.001$ Advanced > Intermediate: $p = 0.036$	Participants with advanced religious knowledge were more likely to verify AI-generated sources, indicating greater critical evaluation.

layers: knowledge integrity (what AI generates and how reliable it is), reasoning capability (how AI interprets and adapts responses), human interaction (how users engage with and rely on AI), and institutional impact (how AI affects authority, governance, and decision structures).

Specifically, the taxonomy is structured into: 1) Epistemic and Knowledge Integrity Risks, which capture concerns related to correctness, evidence, and reliability of AI-generated outputs; 2) Contextual and Reasoning Limitations, which reflect the inability of AI systems to incorporate situational and case-specific nuances; 3) Human–AI Interaction Risks, which describe how users engage with and are influenced by AI systems; and 4) Socio-Technical and Institutional Risks, which capture the broader implications of AI use on authority structures and governance.

This structured approach enables a systematic understanding of participants' perceptions, moving beyond isolated concerns toward a layered interpretation of risks associated with AI-generated guidance.

A. Perceived Risks in AI-Generated Religious Guidance

The analysis of participants' responses revealed a set of distinct but interrelated risk categories associated with the use of AI for generating religious guidance. These risks reflect concerns not only about the technical behaviour of AI systems, but also about their epistemic validity, interaction dynamics, and broader institutional implications.

Table XI presents the finalized taxonomy of perceived risks, organized across the four macro-parameters described above, along with representative quotes from participants.

The identified risks demonstrate that participants' concerns extend beyond isolated technical issues and instead reflect a multi-layered understanding of AI systems operating within a socio-technical context.

TABLE XI. TAXONOMY OF PERCEIVED RISKS WITH REPRESENTATIVE QUOTES

Macro	Parameter	Interpretation	Representative Quotes
Epistemic & Knowledge Integrity	Epistemic Unreliability	Responses may lack correctness, clarity, and evidential strength.	- Reliance on weak evidence - Low accuracy and limited domain knowledge - Answers may contradict established knowledge
	Fabrication & Hallucination	Systems may generate fabricated or non-existent information with high confidence.	- Produces hallucinated information - Cites references that do not exist - Generates incorrect information
	Lack of Authoritative Grounding	Outputs are not anchored in trusted or recognized sources.	- Not connected to recognized scholars - Not based on a trusted source
	Doctrinal Inconsistency	Systems may mix interpretive frameworks without clear methodology.	- Mixing beliefs and sects - Mixing correct and incorrect positions - Manipulation of interpretations
Contextual & Reasoning Limitations	Contextual Insensitivity	Systems fail to account for individual circumstances and situational nuance.	- Cannot adapt rulings to individual situations - Each individual requires a specific ruling
Human–AI Interaction Risks	Preference Alignment Bias	Outputs may align with user expectations rather than truth (sycophancy).	- Distorts answers to satisfy the user - Provides answers at the expense of truth
	Overdependence on AI	Users may rely excessively on AI without verification or independent reasoning.	- Complete reliance on AI - Not using human reasoning or verification
Socio-Technical & Institutional Risks	Displacement of Domain Authority	AI may weaken reliance on domain experts and alter authority structures.	- Abandoning domain experts - Reducing the authority of experts

At the epistemic level (Epistemic and Knowledge Integrity), participants emphasize the importance of correctness, evidence, and source credibility, highlighting concerns about unreliable outputs, fabricated information, and weak grounding in authoritative knowledge. These concerns align closely with the quantitative results, which indicate low trust and high perceived risk.

At the reasoning level (Contextual and Reasoning Limitations), participants identify limitations in the ability of AI systems to handle context-dependent scenarios. This reflects a recognition that certain forms of decision-making require not only information retrieval but also contextual interpretation and situational judgment.

At the interaction level (Human–AI Interaction Risks), concerns shift toward how users engage with AI systems. Participants highlight the risk of overdependence, as well as the tendency of AI systems to produce responses that align with

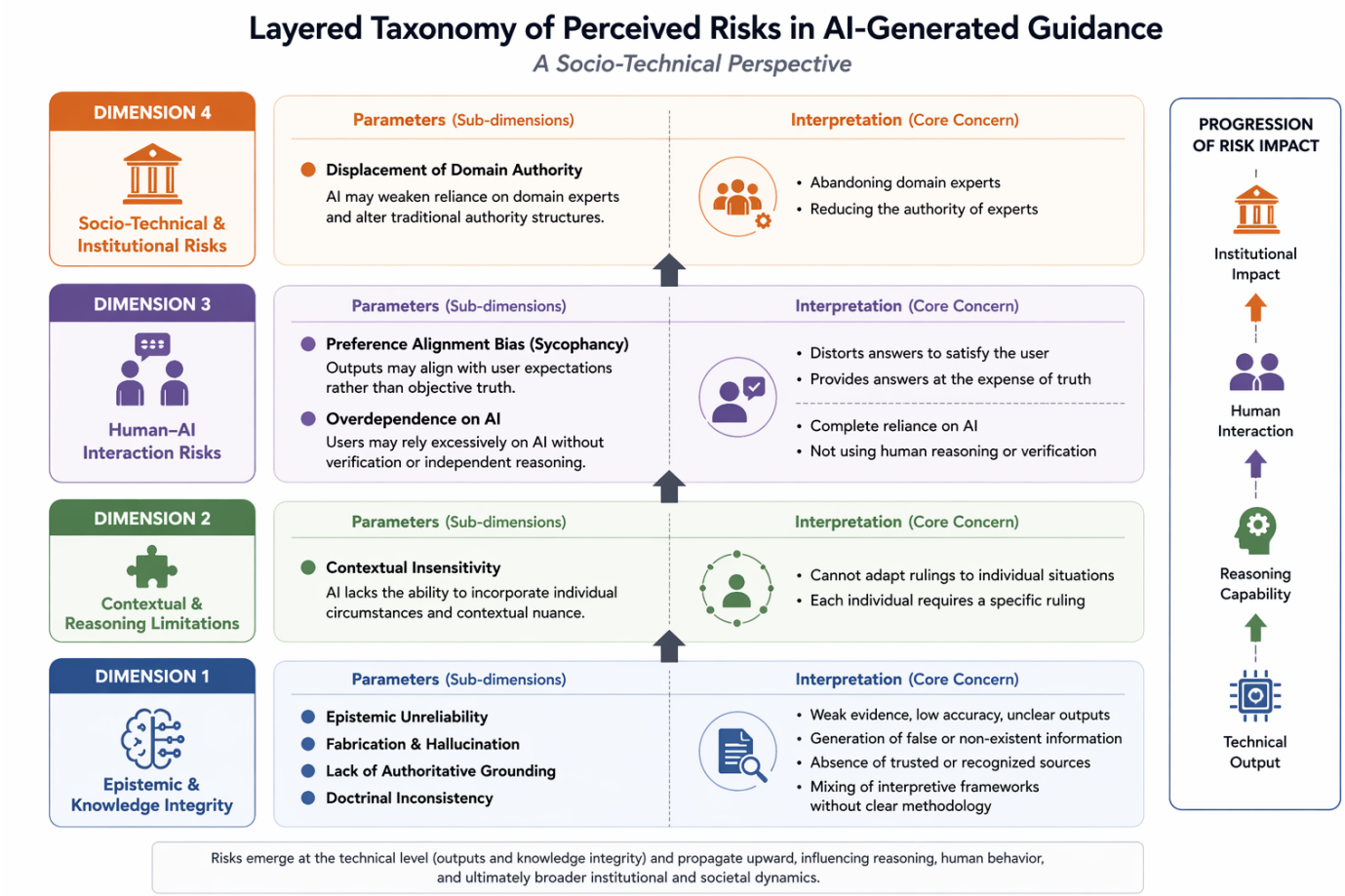


Fig. 9. Framework and layered taxonomy of perceived risks in AI-generated guidance. The figure organizes risks into four macro-parameters: Epistemic and Knowledge Integrity, Contextual and Reasoning Limitations, Human–AI Interaction Risks, and Socio-Technical and Institutional Risks, showing how risks evolve from AI output-level issues to reasoning, interaction, and institutional impacts.

user expectations, potentially reinforcing incorrect beliefs. This behavior aligns with emerging concerns around AI sycophancy, where systems adapt outputs to satisfy users rather than reflect accurate or evidence-based information. This is consistent with observed behavioural patterns in the quantitative analysis, where moderate usage persists despite low trust.

Finally, at the socio-technical level (Socio-Technical and Institutional Risks), participants express concerns about the broader implications of AI adoption, particularly in relation to established authority structures. The potential displacement of domain experts reflects a deeper concern about legitimacy, accountability, and the appropriate role of AI in high-stakes contexts.

These results collectively suggest that participants do not evaluate AI solely based on performance, but rather through a combination of epistemic, behavioural, and institutional considerations. This multi-dimensional perspective is critical for understanding how AI systems are perceived in domains where accuracy, authority, and contextual reasoning are essential.

Fig. 9 presents a visual representation of the framework and layered taxonomy of perceived risks identified in this research. The figure complements Table XI by organizing the parameters into four macro-parameters that reflect distinct but interrelated layers of a socio-technical system. Specifically, the taxonomy illustrates how risks originate at the epistemic and knowledge integrity level, propagate through contextual and reasoning limitations, influence human–AI interaction dynamics, and ultimately extend to broader socio-technical and institutional impacts. This layered structure highlights how lower-level technical limitations in AI outputs can propagate through reasoning and user interaction, ultimately shaping broader institutional and governance outcomes.

B. Governance Expectations for AI Systems

In addition to identifying risks, participants articulated a set of expectations regarding how AI systems should be designed, deployed, and governed in order to be considered acceptable and trustworthy. These expectations reflect a constructive perspective, indicating that users do not reject AI outright but instead advocate for its controlled and responsible use. These expectations also implicitly relate to privacy concerns identified in the quantitative analysis, particularly in scenarios where users may prefer AI-mediated interactions for sensitive inquiries, while remaining uncertain about data handling and confidentiality.

The analysis revealed two primary parameters of gover-

nance expectations: 1) Transparency and verification mechanisms, and 2) Human oversight and institutional governance.

The first category emphasizes the need for AI-generated responses to be supported by clear, verifiable, and authoritative sources. Participants consistently highlighted the importance of transparency in how responses are generated, particularly the inclusion of reliable references that can be independently verified. This expectation directly corresponds to concerns related to epistemic unreliability and lack of authoritative grounding, suggesting that users seek mechanisms that enhance trust through evidence and traceability.

The second category focuses on the role of human expertise and institutional oversight. Participants emphasized that AI systems should not operate independently in high-stakes domains, but rather under the supervision of qualified experts and recognized institutions. This reflects a strong preference for maintaining human accountability and preserving established authority structures, particularly in contexts where interpretive judgment is required.

These governance expectations demonstrate that participants envision a hybrid model of AI use, in which AI systems function as supportive tools rather than autonomous decision-makers. The alignment between identified risks and proposed governance mechanisms highlights a coherent user perspective: the same concerns that reduce trust in AI also inform the conditions under which AI may become acceptable.

VI. Discussion

This research provides a comprehensive understanding of how individuals perceive the use of AI systems in high-stakes, interpretive domains by integrating quantitative and qualitative analyses, with a particular focus on religious inquiry in Saudi Arabia. The results reveal a consistent and nuanced pattern: participants engage with AI-generated guidance in a limited and cautious manner, characterized by moderate usage, low trust, and high perceived risk, while still recognizing the potential utility of these systems under constrained conditions.

The quantitative results establish the structural patterns underlying this behaviour. Participants reported only occasional use of AI for religious inquiries, despite widespread exposure to such tools. Trust emerged as a central determinant of behavioural intention, with higher trust associated with increased willingness to use AI and, in some cases, to follow its outputs even when they conflict with trusted human experts. However, overall trust levels remained low, while perceived risk remained high. This combination indicates not rejection, but selective and conditional engagement. Participants do not treat AI as an authoritative source, but neither do they disregard it entirely. Instead, they adopt a cautious usage model in which AI is consulted but not relied upon without verification. This is further reinforced by strong preferences for source credibility, where the majority of participants actively verify AI-generated outputs and express a clear expectation that responses be linked to recognized scholarly authorities.

These patterns are further shaped by demographic and knowledge-based factors. Younger participants demonstrate greater openness to AI use, particularly in sensitive contexts, suggesting that accessibility and privacy-related affordances may lower barriers to engagement. In contrast, participants with stronger domain knowledge exhibit lower trust, higher perceived risk, and more rigorous verification behaviour. This indicates that familiarity with the domain introduces critical evaluation mechanisms that limit reliance on AI-generated outputs. Together, these results highlight that engagement with AI is not uniform, but mediated by both technological familiarity and epistemic grounding.

While the quantitative results describe these behavioural patterns, the qualitative analysis explains their underlying structure. The taxonomy of perceived risks reveals that participants evaluate AI systems across multiple interconnected dimensions rather than isolated concerns. At the epistemic level, concerns focus on correctness, evidence, and reliability, including risks of weak grounding, fabricated information, and inconsistent interpretations. At the reasoning level, participants highlight limitations in contextual understanding, recognizing that AI systems cannot adequately account for situational nuance or case-specific factors. At the interaction level, concerns shift toward behavioural dynamics, including overdependence and preference alignment, where AI systems may produce responses that align with user expectations rather than objective truth, reflecting emerging concerns around AI sycophancy. At the socio-technical level, participants express concern about the broader implications of AI adoption, particularly the potential displacement of domain authority and the erosion of established structures of expertise.

The integration of these results represents a key contribution of this work. Rather than treating trust, risk, and behaviour as independent variables, the results demonstrate that they are structurally linked. Low trust and high perceived risk observed in the quantitative analysis are directly explained by epistemic and reasoning concerns identified in the qualitative results. Similarly, observed verification behaviours correspond to concerns about credibility and lack of authoritative grounding, while conditional acceptance of AI aligns with expectations for human oversight and governance. This layered interpretation shows that user perception of AI is not driven solely by system performance, but by how technical limitations propagate through reasoning, interaction, and institutional contexts.

These results should be understood within a broader transformation in how knowledge and authority are mediated in digital environments. Existing literature suggests that AI in such domains does not merely expand access to information but reshapes how authority is constructed, circulated, and evaluated [29]. Prior work has shown that users increasingly question the credibility, contextual relevance, and institutional grounding of mediated guidance, particularly in domains where expert interpretation and contextual reasoning are essential [30]. The present research extends this literature by providing empirical evidence from Saudi Arabia on how users actively negotiate this shift in practice: while AI systems improve accessibility and immediacy, their acceptance remains conditional on trust, verification, and alignment with recognized sources of authority.

The results also align with and extend recent work on AI-generated religious content and Islamic knowledge systems. Alam et al. [11] demonstrate, in a study of Muslim American users, that users may prefer AI-generated religious answers even when expert evaluation identifies significant quality con-

cerns, highlighting the gap between perceived usefulness and actual reliability. The present results extend this understanding by examining similar dynamics within Saudi Arabia, where expectations of authority, trust, and verification are shaped by a distinct religious, cultural, and institutional context. Similarly, Priantina et al. [12] show that while AI can improve efficiency in religious ruling related processes, it raises important concerns about automating scholarly reasoning and weakening human oversight. Bhatia et al. further identify broader challenges in AI-based Islamic knowledge systems, including trustworthiness, evaluation standards, pluralism, and governance gaps [13]. This research contributes to these discussions by shifting the focus from system performance to user perception, demonstrating how trust, risk, privacy, credibility, and authority are evaluated in real-world use rather than controlled system settings.

More broadly, the results connect with established challenges in large language models and generative AI. Participants' concerns about inaccurate or misleading outputs correspond directly to the well-documented problem of hallucination, where systems generate fluent but factually incorrect or unsupported information [4], [27]. In high-stakes contexts, such outputs may be interpreted as authoritative despite lacking reliable grounding. Similarly, concerns about preference alignment reflect emerging work on AI persuasion and safety, where generative systems may shape user beliefs and decisions rather than simply provide neutral information [5], [31]. These risks are consistent with broader taxonomies of language-model harms, including misinformation, bias, privacy risks, and interaction-level failures [32], [33].

The results also resonate with literature on trust, explanation, and human–AI interaction. Research on trust in automation shows that achieving appropriate reliance is inherently difficult, as users may either over-rely on automated systems or reject them entirely depending on perceived reliability [34]. In contrast, this research identifies a more nuanced behavioral pattern characterized by cautious engagement and active verification, suggesting the emergence of calibrated trust in high-stakes contexts. At the same time, explainability research indicates that explanations can improve user understanding but may also increase acceptance of incorrect outputs when users lack sufficient expertise to critically evaluate them [35], [36]. This highlights a critical implication: transparency and explainability are necessary but not sufficient. In interpretive domains, they must be complemented by expert validation, institutional oversight, and mechanisms for accountability.

From a theoretical perspective, this research contributes a socio-technical model of AI risk perception in which concerns emerge across interconnected layers rather than isolated system failures. By organizing perceived risks into epistemic, reasoning, interaction, and institutional dimensions, the proposed taxonomy provides a structured framework for understanding how technical limitations in AI systems propagate into behavioral and governance challenges. This extends existing work on foundation models and AI risk by grounding abstract categories in empirically observed user perceptions and linking system-level behaviour with real-world interpretation and use [32], [33], [37]. In doing so, the research bridges technical AI research with human-centered and governance-oriented perspectives on trust, authority, and decision-making.

The results also have important implications for AI governance and system design [38], [39]. Participants consistently emphasized the need for transparency, particularly through the provision of verifiable and authoritative sources. However, the results suggest that transparency alone is insufficient in high-stakes interpretive contexts. Users not only expect explanations, but also require mechanisms to assess their validity, especially when they lack the expertise to independently verify the information provided. This reinforces the need for governance frameworks that combine explainability with validation, accountability, and oversight [35], [36], [37]. In addition, participants consistently rejected the idea of AI systems functioning as autonomous authorities. Instead, they expressed a strong preference for models in which AI operates under human supervision, particularly by domain experts. This aligns with broader risk-based governance approaches, where human-in-the-loop mechanisms are considered essential for high-risk applications [34], [37].

Importantly, the results highlight a tension between accessibility and authority. While AI systems increase access to information and provide immediate responses, they do not inherently carry the legitimacy associated with expert judgment. Participants' emphasis on verification and reliance on trusted sources indicates that authority remains a central condition for acceptance, even in the presence of highly accessible AI tools. Rather than replacing established structures of expertise, AI systems are therefore more likely to be accepted as assistive tools that complement, rather than substitute, human authority. This reflects a broader pattern observed in other high-stakes domains, where users value the efficiency and clarity of AI systems but continue to depend on expert validation for final decisions [40].

From a practical perspective, the results suggest that AI systems in high-stakes advisory domains require governance models that combine verifiable source grounding, human oversight, and institutional accountability. These expectations align with broader concerns about foundation models, where scale, opacity, and downstream deployment amplify risks across domains [37]. They also reflect patterns observed in other domains such as healthcare, where users value AI-generated responses for accessibility and clarity but continue to rely on human experts for validation and safe decision-making [40]. Together, these results point toward a hybrid governance model that supports human autonomy, enables calibrated trust, and preserves the role of domain expertise.

Finally, while this research is situated within a specific cultural and religious context, the underlying dynamics are likely to extend to other high-stakes interpretive domains, including healthcare, legal advisory, education, and policy decision-making, where similar concerns around trust, expertise, and decision-making have been observed [40], [1], [2]. In these contexts, trust, privacy, contextual reasoning, authority, and institutional legitimacy remain central to responsible AI deployment. Accordingly, the proposed framework should be understood as a structured analytical foundation for examining how technical, behavioural, and institutional risks interact in high-stakes interpretive settings, rather than as a universally validated model across all domains. The research, therefore, contributes to a broader shift in AI research from system-centric performance toward human-centred and governance-

oriented understanding of AI in practice.

VII. Conclusion

This research provides a comprehensive examination of how individuals perceive the use of AI systems in high-stakes interpretive domains, using the context of religious inquiries in Saudi Arabia as an empirical setting. By integrating quantitative and qualitative analyses, the results reveal a consistent pattern characterized by moderate engagement with AI tools, low levels of trust, and heightened perceptions of risk. Users demonstrate a cautious and conditional approach, treating AI-generated responses as supportive resources rather than authoritative sources.

The results highlight that trust, perceived risk, and source credibility are central determinants of user behaviour. Participants consistently emphasized the importance of verifiable and authoritative grounding, frequently engaging in source verification and expressing strong preference for responses linked to recognized experts. At the same time, reluctance to rely on AI for sensitive or complex inquiries underscores the context-dependent nature of AI acceptance. Demographic factors, particularly age and domain knowledge, further shape how individuals interpret, evaluate, and interact with AI-generated outputs.

Beyond descriptive insights, this research advances understanding by introducing a structured, multi-layered taxonomy of perceived risks, spanning epistemic, reasoning, interactional, and institutional dimensions. This framework demonstrates that concerns are not limited to technical limitations, but extend to broader issues of authority, accountability, and governance. The alignment between identified risks and user expectations further reveals that acceptance of AI is contingent on the presence of transparency, human oversight, and institutional anchoring.

These results carry important implications for the design and governance of AI systems in high-stakes contexts. Enhancing transparency through verifiable sourcing, incorporating human-in-the-loop mechanisms, and establishing clear governance boundaries are essential for building trustworthy and socially aligned AI systems. Rather than replacing established structures of expertise, AI systems are more likely to be accepted when positioned as assistive tools operating within well-defined ethical and institutional frameworks.

While this research is grounded in a specific cultural and religious context, the underlying dynamics are likely to extend to other domains where interpretive judgment, trust, and accountability are critical. At the same time, the participant distribution was skewed toward younger and relatively educated respondents within a single national setting, which may influence the broader generalizability of the reported perceptions and behavioral patterns. This positions the work within a broader discourse on responsible AI, highlighting the need to align technological capabilities with human expectations, domain expertise, and societal values, while also motivating future cross-cultural and demographically diverse investigations.

Future research can build on these results by exploring cross-cultural variations, longitudinal shifts in trust, and experimental evaluations of AI systems integrated with expert validation mechanisms. Additional directions include real-world system evaluations, adversarial robustness testing of AI-generated guidance, and longitudinal trust calibration studies examining how user perceptions evolve through repeated interaction with AI systems over time.

To conclude, AI systems hold significant potential to enhance access to complex knowledge domains. However, their acceptance depends not only on technical performance, but also on their ability to operate within trusted epistemic and institutional structures. The challenge, therefore, is not merely to improve AI systems, but to design them in ways that respect, support, and reinforce the foundations of human expertise and judgment.

Acknowledgment

This study is derived from a research grant funded by the Research, Development, and Innovation Authority (RDIA), Kingdom of Saudi Arabia, with grant number 12615-iu-2023-IU-R-2-1-EI-.

References

[1] S. Shekar, P. Pataranutaporn, C. Sarabu, G. A. Cecchi, and P. Maes, "People over-trust ai-generated medical responses and view them to be as valid as doctors, despite low accuracy," *arXiv preprint arXiv:2408.15266*, Aug 2024.

[2] E. Schneiders, T. Seabrooke, J. Krook, R. Hyde, N. Leesakul, J. Clos, and J. E. Fischer, "Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm." Association for Computing Machinery, 4 2025. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3706598.3713470

[3] T. Seabrooke, E. Schneiders, L. Dowthwaite, J. Krook, N. Leesakul, J. Clos, H. Maior, and J. Fischer, "A survey of lay people's willingness to generate legal advice using large language models (llms)," *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, 9 2024.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[5] H. Kong, "Persuasion and safety in the era of generative ai," in *WebSci PhD Symposium*, 2025.

[6] C. Cuskley, R. Woods, and M. Flaherty, "The limitations of large language models for understanding human language and cognition," *Open Mind*, vol. 8, pp. 1058–1083, 08 2024. [Online]. Available: https://doi.org/10.1162/opmi_a_00160

[7] N. Yousif, "Parents of teenager who took his own life sue openai," *BBC News*, 9 2025. [Online]. Available: https://www.bbc.com/news/articles/cgerwp7rdlvo

[8] R. Alsaigh, R. Mehmood, I. Katib, and T. Yigitcanlar, "Governing ai in society: Explainable analysis of research using the pearl methodology, the frame-ai framework, and regulatory alignment gaps," *SSRN Electronic Journal*, Dec. 2025. [Online]. Available: https://ssrn.com/abstract=6009614

[9] R. Alsaigh, R. Mehmood, I. Katib, X. Liang, A. Alshanqiti, J. M. Corchado, and S. See, "Harmonizing AI governance regulations and neuroinformatics: perspectives on privacy and data sharing," *Frontiers in Neuroinformatics*, vol. Volume 18 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2024.1472653

[10] S. A. Al-Tayyar, "Fatwa and its importance." [Online]. Available: https://draltayyar.com/books/7963/

[11] S. M. Alam, M. Abdulhai, and N. Salehi, "Blind faith? user preference and expert assessment of ai-generated religious content," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2025, pp. 1–29.

[12] A. Priantina, M. M. Uula, Aufa, and E. Herindara, "Ai in fatwa formulation: Transforming sharia-compliant finance," *Journal of Central Banking Law and Institutions*, vol. 4, no. 3, pp. 595–634, 2025.

[13] G. Bhatia, H. Mubarak, M. Hawasly, M. Jarrar, G. Mikros, F. Zaraket, M. Alhirthani, M. Al-Khatib, L. Cochrane, K. Darwish, R. Yahiaoui, and F. Alam, "Advances in ai systems on islamic knowledge capabilities: A critical survey," *arXiv preprint*, 2026, available at: https://gagan3012.github.io/islamic-knowledge-survey/.

[14] O. Alyemny, H. Al-Khalifa, and A. Mirza, "A data-driven exploration of a new islamic fatwas dataset for arabic nlp tasks," *Data 2023, Vol. 8, Page 155*, vol. 8, p. 155, 10 2023. [Online]. Available: https://www.mdpi.com/2306-5729/8/10/155/htmhttps://www.mdpi.com/2306-5729/8/10/155

[15] A. Y. Alan, E. Karaarslan, and Ömer Aydin, "A rag-based question answering system proposal for understanding islam: Mufassirqas llm," *MufassirQAS. Turkish Journal of Engineering*, vol. 9, pp. 544–559, 1 2024. [Online]. Available: https://arxiv.org/pdf/2401.15378

[16] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag," *Neural Computing and Applications*, vol. 37, pp. 20 957–20 982, 9 2025. [Online]. Available: https://link.springer.com/article/10.1007/s00521-025-11229-y

[17] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[19] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[20] "General presidency of scholarly research and ifta." [Online]. Available: https://alifta.gov.sa/home

[21] "Bin othaimeen." [Online]. Available: https://binothaimeen.net/site

[22] "Fatawa pedia." [Online]. Available: https://fatawapedia.com/

[23] H. Huang *et al.*, "Acegpt, localizing large language models in arabic," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 1, pp. 8132–8156, 9 2023. [Online]. Available: https://arxiv.org/pdf/2309.12053

[24] P. Georgiev *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 3 2024. [Online]. Available: https://arxiv.org/pdf/2403.05530

[25] "Silma ai: The leading arabic ai & llm technology provider." [Online]. Available: https://silma.ai/

[26] B. Wen *et al.*, "Know your limits: A survey of abstention in large language models." [Online]. Available: https://doi.org/10.1162/tacl

[27] F. Atif, N. Askarbekuly, K. Darwish, and M. Choudhury, "Sacred or synthetic? evaluating llm reliability and abstention for religious questions," 8 2025. [Online]. Available: https://arxiv.org/pdf/2508.08287

[28] G. Inoue, B. Alhafni, R. Baly, F. Zaraket, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 92–104.

[29] F. A. Atallah, "Digital mediation and fatwa authority in contemporary islam: A critical islamic legal and media-theoretical framework," *Religions*, vol. 17, no. 350, 2026.

[30] S. Whyte, "Are fatwas dispensable? examining the contemporary relevance and authority of fatwas in australia," *Oxford Journal of Law and Religion*, vol. 11, pp. 314–342, 2022.

[31] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[32] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[33] L. Weidinger *et al.*, "Taxonomy of risks posed by language models," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[34] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[36] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[37] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx *et al.*, "On the opportunities and risks of foundation models," Stanford Center for Research on Foundation Models (CRFM), Tech. Rep., 2021, arXiv:2108.07258.

[38] R. Alsaigh, R. Mehmood, and I. Katib, "AI explainability and governance in smart energy systems: A review," *Frontiers in Energy Research*, vol. Volume 11 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2023.1071291

[39] R. Alsaigh, R. Mehmood, I. Katib, A. A. Almuzaini, S. S. Albouq, and S. Alshmrany, "Evidence-Driven AI Governance for Healthcare: A PEARL-PATHWAY Analysis of Madinah," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 4, 2026. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=17&Issue=4&Code=IJACSA&SerialNo=XX

[40] S. K. Beale, N. Cohen, B. Secheli, D. McIntire, and K. A. Kho, "Comparing physician and artificial intelligence chatbot responses to posthysterectomy questions posted to a public social media forum," *AJOG Global Reports*, vol. 5, p. 100553, 2025.

D-HAN-Net: A Hybrid Dual-Stream Architecture for Corporate Bankruptcy Prediction via Multimodal Fusion

Md. Abul Kalam Azad[1], Abdul Kadar Muhammad Masum[2*], Najmus Saadat[3], Esrat Jahan[4],
Ramona Birău[5], Virgil Popescu[6], Iuliana Carmen Bărbăcioru[7], Ștefan Mărgăritescu[8]
Dept. of Business and Technology Management, Islamic University of Technology, Gazipur 1704, Bangladesh[1]
Dept. of Computer Science and Engineering, Southeast University, Dhaka 1208, Bangladesh[2]
Computational Intelligence Lab, Southeast University, Dhaka 1207, Bangladesh[3,4]
"Eugeniu Carada" Doctoral School of Economic Sciences, University of Craiova, Craiova, Romania[5,8]
Faculty of Economic Science, Constantin Brâncusi University of Târgu Jiu, 210185 Târgu Jiu, Romania[5]
Faculty of Economics and Business Administration, University of Craiova, Craiova, Romania[6]
Faculty of Engineering, University of Târgu Jiu, Gorj county, Romania[7]

***Abstract*—Early detection of corporate bankruptcy is essential for maintaining financial market stability and reducing systemic risk. However, existing predictive models often struggle under conditions of extreme class imbalance. Traditional approaches either analyze financial ratios or textual disclosures in isolation, while conventional cross-modal fusion strategies tend to dilute rare distress signals when integrating structured accounting metrics with qualitative management narratives. Consequently, subtle indicators of impending corporate failure are frequently overshadowed by dominant non-distress patterns, limiting the effectiveness of existing predictive systems. To bridge this methodological limitation, this study proposes D-HAN-Net, a hybrid dual-stream deep learning architecture. This framework is intended to address minority class suppression by dynamically balancing structured financial ratios with unstructured textual disclosures. Specifically, the model processes numerical indicators through a gated residual network to capture complex patterns. Simultaneously, it extracts semantic cues from corporate reports via a FinBERT-driven bidirectional GRU. These dual modalities are then aligned using a learnable cross-modal attention fusion gate, jointly optimized with focal loss. Experimental evaluations on a comprehensive multimodal dataset, utilizing stratified splits demonstrate that D-HAN-Net significantly outperforms state-of-the-art baselines, achieving a predictive accuracy of 94.00%, an F1-score of 88.00%, and an AUC of 0.9734. Practically, this framework equips investors, financial institutions, and regulatory authorities with a decisive early warning system. It enables proactive risk management by detecting subtle distress signals before corporate failure becomes irreversible. Furthermore, extensive stability testing and ablation analysis confirm that the model's superior predictive reliability is highly robust against sampling uncertainty, fundamentally relying on the synergistic integration of all its architectural modules.**

***Keywords*—Bankruptcy prediction; multimodal deep learning; corporate disclosures; cross-modal fusion; class imbalance; early warning system**

I. Introduction

Corporate bankruptcy prediction has become a central research problem in financial analytics, as the early identification of financially distressed firms is very critical for investors, creditors, regulators and other market participants. Accurate prediction models can minimize the potential losses, facilitate the timely interventions, and enhance the distribution of financial resources across the capital markets. The number of corporate bankruptcies is also generally believed to be a signal of overall economic health and financial stability, with recent studies like Ji et al. [1] emphasizing that such models serve as critical early warning systems for stakeholders. For this reason, bankruptcy prediction is still an important part of financial risk assessment and corporate decision support. Traditionally, the goal of bankruptcy prediction models has been to assess the financial situation of a firm and forecast the probability of failure before distress is irreversible, a concept further advanced by Lin et al. [2] through modern explainable machine learning integrations. Most previous research tackled this task by means of the financial statement analysis by using financial statements accounting based indicators created from balance sheets, income statements and cash flow statements. These structured numerical variables have long been used as the basis for both statistical and machine learning based bankruptcy prediction models. Prior work indicates that machine learning can surpass traditional statistical methods by capturing complex nonlinear relationships between financial indicators. Recent explainable and hybrid approaches proposed by Fasano et al. [32] and Ainan et al. [34] further demonstrate the growing effectiveness of intelligent predictive systems in bankruptcy analysis. For instance, Liu et al. [3] report that machine learning models optimized with advanced algorithms are able to perform prediction much better, and Jiang et al. [4] highlight the growing importance of such intelligent computational approaches to financial distress studies. However, despite this progress, bankruptcy prediction remains challenging due to the rarity, heterogeneity, and complexity of real-world corporate failure patterns. Corporate bankruptcy prediction can be viewed as an imbalanced multimodal financial risk classification problem in which structured accounting variables and unstructured narrative disclosures must be jointly modeled.

A key issue in this literature is the extreme class imbalance that is present in most bankruptcy data. In practice, bankrupt companies are a small fraction of all companies, which often

*Corresponding author

leads prediction models to be biased towards the majority class of healthy companies and perform poorly on the minority class of distressed companies. Lombardo et al. [5] point out the practical difficulty of bankruptcy prediction under highly imbalanced settings, whereas Ranjbaran et al. [6] demonstrate that augmentation based strategies might lead to better minority detection, but at the same time can lead to instability and increased false positive rates. Beyond the imbalance problem, another important limitation is the high level of dependence of many prediction systems on structured financial ratios alone. Although such variables capture quantitative aspects of profitability, liquidity, leverage and solvency, the development of financial distress rarely occurs through numerical deterioration only. In real corporate situations, the warning signs of distress may often appear in managerially talk, auditor's concerns and forward looking uncertainty before it may be fully reflected in the accounting results. Corporate narrative disclosures, particularly the Management Discussion and Analysis section and auditor reports, therefore, lend a further dimension of risk relevant information. Mai et al.[7] demonstrate the use of textual disclosures from corporate filings to further improve bankruptcy prediction since they capture the semantic patterns that are unavailable in structured data. Related accounting research also lends credence to the informational value of narrative reporting. Mayew et al. [8] find that Management Discussion and Analysis disclosures to deliver incremental evidence about the firm's ability to continue as a going concern. Li et al. [9] show that readability in financial reporting is associated with earnings persistence and transparency while Loughran and McDonald [10] demonstrate that linguistic tone in corporate filings carries risk related information beyond conventional accounting measures. From a theoretical perspective these results imply that numerical indicators and textual disclosures are complementary as opposed to competing information channels.Financial ratios summarize realized economic outcomes, while narrative disclosures communicate managerial interpretation, uncertainty and emerging risks that may not yet be fully reflected in accounting variables. This view is consistent with both information asymmetry as well as signaling theories of financial reporting, where managers convey firm conditions in the form of both formal accounting results and qualitative disclosures. Yet most bankruptcy prediction studies still consider these two modalities separately. Although recent multimodal frameworks such as Mancisidor and Aas [37] indicate that combining textual and numerical information can substantially improve predictive capability. As a result, existing approaches often fail to effectively capture cross-modal interactions between quantitative financial deterioration and qualitative narrative distress signals.

This limitation leads to an obvious research gap and directly leads in to the present study. Despite the recent progress in deep learning, specifically transformer-based language models like BERT by Devlin et al. [11], the representation of long textual documents has been significantly improved. However, bankruptcy prediction research is still missing a unified framework that can effectively integrate structured financial indicators with corporate narrative disclosures under highly imbalanced conditions. Recent evidence from Chen et al. [12] suggests that textual information extracted from the annual reports can improve the predictive performance when integrated into learning-based systems. suggests that textual information extracted from annual reports can improve predictive performance when integrated into learning based systems. Similar findings have also been reported in recent multimodal and hybrid bankruptcy prediction studies such as Arno et al. [36], but the broader challenge of theoretically and computationally combining text and numerical signals is still to be solved. Nevertheless, modeling lengthy financial disclosures remains challenging because important distress-related evidence may be distributed across multiple document segments rather than localized within short textual spans. This encourages the following research question: How can structured financial indicators and narrative corporate disclosures be successfully combined in a unified predictive model in order to better detect corporate financial distress under highly imbalanced bankruptcy datasets?

To answer this question, this study proposes D-HAN-Net, a hybrid dual-stream deep learning architecture that considers the analysis of structured numerical indicators and corporate textual disclosures jointly. The framework uses a gated residual network to process financial variables to model nonlinear dependencies among ratios. It also employs a transformer-based language encoder followed by a hierarchical recurrent aggregation layer to model the text disclosures and retain the contextual semantic structure across long documents. This design also improves computational efficiency and training stability under limited bankruptcy data conditions. A cross-modal fusion mechanism then combines the two modalities and dynamically balances the relative contributions of the two modalities during the prediction process. The main contributions of this study are summarized as follows: First, it constructs a multimodal deep learning architecture for joint modeling of financial ratios and corporate narrative disclosures for bankruptcy prediction. Second, it presents a cross-modal fusion strategy that leverages the complementary nature between numerical and textual signals in an integrated architecture. Third, it illustrates the efficacy and robustness of the proposed framework with extensive empirical analysis of the framework, including baseline comparison, ablation analysis, and cross seed stability analysis.

II. Literature Review

Bankruptcy prediction has long been studied in financial economics, accounting, and data science. Over the years, the methods of research have evolved from old statistical models to new machine learning and deep learning techniques. Financial distress theory and information asymmetry literature imply that corporate distress signals arise from various sources, which includes financial statements and corporate disclosures. Consequently, both financial indicators and narrative information can offer valuable signals for identifying financial difficulty faced by firms in the form of both quantitative financial indicators and qualitative narrative information. This section provides an overview of the theoretical background of bankruptcy prediction, and an overview of previous research works concerning machine learning models, text financial analysis and imbalance aware learning models.

Corporate bankruptcy generally has been explained in terms of theories of financial distress between firm failure and deterioration in financial structure, liquidity, increasing leverage, and falling profitability. Bankruptcy risk, therefore,

becomes visible through signals such as the worsening of liquidity, unstable cash flows, declining profitability and growing pressure to ensure solvency. For this reason, financial indicators based on financial statements have been the main inputs for bankruptcy prediction models for a long time [13]. However, financial distress is not only reflected in the accounting numbers. Disclosure theory implies that managers are communicating risk via narrative disclosures in annual reports and regulatory filings. When firms find themselves under financial stress, discussions among managers usually involve problems with operations, problems with financing, or restructuring activities. Such signals often occur in advance of the time when deterioration is fully reflected in the accounting ratios. Financial language models like FinBERT can capture these contextual cues because they are trained on financial corpora and are able to spot domain specific expressions related to uncertainty and financial instability. Taken together, these views mean that financial distress is multidimensional. Numerical indicators provide information on measured deterioration in firm performance, while narrative disclosures provide managerial perceptions and advance warning signals about financial stress. Integrating financial indicators and FinBERT derived semantic representations, therefore, offers a theoretically grounded basis for enhancing bankruptcy prediction models [14], [15].

Machine learning techniques have been popularly used to predict bankruptcy because of their capability of capturing complex nonlinear relationships between financial variables and bankruptcy outcomes. Studies have been done on the neural network and the sequential learning models to detect the financial distress pattern. Kim et al. showed that the modeling of temporal dependencies between financial indicators is useful for enhancing prediction performance [16]. Hosaka proposed transforming financial ratios into image representations and using convolutional neural networks to identify patterns in structure of financial data [17]. Pellegrino et al. further developed an architecture of multiple heads Long Short Term Memory to learn temporal dynamics from accounting time series data [18]. Ensemble learning methods have also attracted a lot of attention. Recent interpretable ensemble frameworks combining models like LightGBM and CatBoost with SHAP analysis demonstrate significantly more robust and transparent predictions across diverse market conditions [19]. Liang et al. showed that the integration of corporate governance index and financial index can improve the predictive effect and will show the importance of diverse corporate characteristics in the predictive model [20].

Recent researches have been devoted to advanced models that can be applied to high-dimensional financial dataset. Gradient boosting methods like LightGBM have been found to have good predictive power in financial risk modeling [21]. Ensemble approaches have also been shown to have great potential. Entropy based feature transformation strategies which enhances predictive performance in financial risk modeling was introduced by Carta et al. [22], [23]. Stacking based ensemble methods further suggests that a combination of classifiers will yield more stable and reliable results than relying on a single model [24]. Feature selection methods have also been proven to be efficient to identify the most informative indicators to predict bankruptcy [25]. Research and surveys have shown that machine learning and deep learning methods typically outperform conventional statistical models in bankruptcy prediction tasks [26], [27]. Financial distress is communicated not only through numerical ratios but also through managerial narratives in corporate disclosures. Managers often adjust the tone and wording of reports when firms face liquidity pressure, operational challenges, or solvency concerns. As a result, textual disclosures may reveal early signals of distress before these conditions become fully visible in financial statements. Transformer based language models such as BERT can capture contextual semantic relationships in textual data [11]. In financial applications, domain-adapted models such as FinBERT, are particularly effective at identifying linguistic patterns related to uncertainty, litigation risk, and financial instability. Chen et al. demonstrated that incorporating textual features from annual reports significantly improves bankruptcy prediction performance [12]. Lombardo et al. further showed that deep language models can effectively analyze complex corporate disclosure documents [28].

One of the biggest challenges in predicting bankruptcy lies in the extreme class imbalance with financial datasets, with bankrupt firms only making up a small percentage of observations. This imbalance will often lead to predictive models giving preference to the majority class, and making it difficult for them to accurately determine correct cases of bankruptcy. To overcome this problem, imbalance aware learning strategies have been proposed. Recently, Papík demonstrated the significant impact of advanced data resampling approaches, such as SMOTE and random oversampling, on improving minority class detection for imbalanced financial datasets [29]. Kubat and Matwin have introduced one sided selection techniques that reduce the imbalance by eliminating redundant majority class samples from the training data [30]. These approaches have become important parts of modern financial risk predicting systems. Despite some significant progress, there are still some limitations in the literature. Many of the researchers mainly focus on financial indicators and pay little attention to qualitative signals in corporate disclosures. Studies that look at the analysis of textual information tend to look at the narrative signals independent of financial variables leaving the interaction between these two parameters insufficiently explored. Furthermore, many models work such that they fit within a single modality framework and cannot, therefore, capture complementary distress signals from heterogeneous data sources.

To overcome these limitations, this research study proposes D-HAN-Net, a dual stream hybrid deep learning architecture that jointly analyzes structured financial data and semantic information from the corporate disclosures. The model uses separate encoders for numerical and textual modalities and uses an adaptive cross-modal fusion mechanism to dynamically balance contributions of the two modalities during the prediction. By incorporating heterogeneous financial signals into an integrated framework, the proposed approach is expected to enhance the bankruptcy prediction performance and result in a more robust bankruptcy early warning system under class imbalanced conditions.

III. Methodology

This section presents the architecture design of the proposed D-HAN-Net which is schematically shown in Fig. 1.

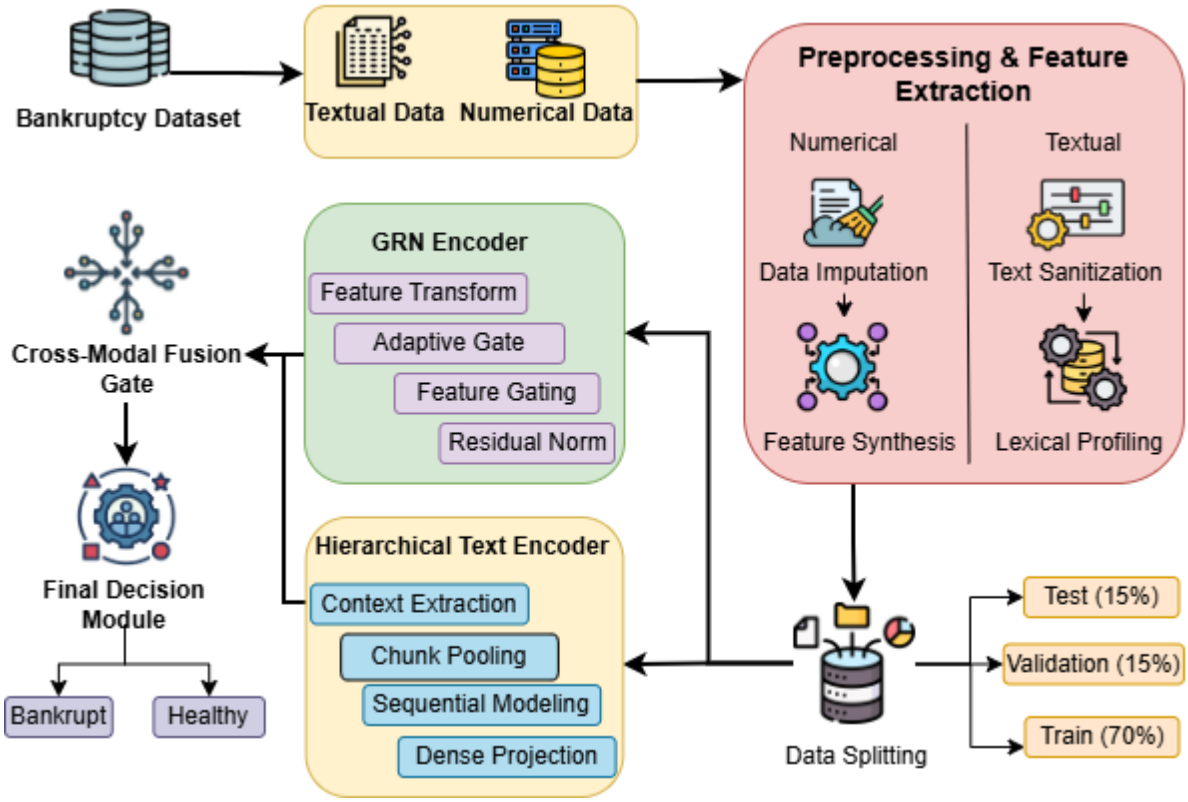


Fig. 1. End-to-end architecture of the proposed dual modality learning framework for bankruptcy prediction.

The framework consists of a combination of multimodal cohort construction, disclosure-driven risk feature engineering and a hybrid neural architecture with gated numerical encoding, hierarchical document encoding and cross-modal fusion to predict next year bankruptcy.

A. Dataset Description

This study is based on the publicly available Figshare dataset of multimodal bankruptcies [31]. In each instance there are MD&A text, audit opinion text, structured numerical attributes and the next year bankruptcy label binary (indexed by UID). The circumstances are not in our favor since bankrupts are few and financial filings are long. Therefore, the decision relevant signals are preserved by imbalance aware sampling and hierarchical text processing. Potential selection bias could also arise due to the fact that financially distressed firms have larger management reports and larger auditor opinions. Distressed companies often offer specific explanations about the reasons for liquidity situations, operational uncertainty, debt commitments, or restructuring efforts. Consequently the intensity of narrative may be partly correlated with the risk of bankruptcy. To reduce this concern, the proposed framework is built on the basis of contextual semantic representations that are derived from segmented disclosures rather than using raw document length as a direct predictor. Long disclosures are segmented into multiple chunks to reduce excessive truncation while preserving contextual semantic information across lengthy financial reports.

B. Data Preprocessing

The pipeline begins by reading both of the raw sources in chunks of size 5000 avoiding memory overflow and then aligning them by UID with an inner merge so that every retained record contains both modalities. In order to minimize the imbalance while preserving all the positive samples, all bankrupt UIDs are kept. To address the serious class imbalance typical of bankruptcy data, all bankrupt companies are kept but a sample of healthy companies is randomly selected in a ratio of around 5:1. This sampling strategy mitigates the extreme imbalance without having too few observations from the majority class for stable model training. The last cohort is divided by stratified sampling into train 70%, validation 15% and test 15%, which makes the class ratios stable between partitions and prevents the statistical instability that time-based splits would cause due to the extreme rarity of bankruptcies. The resulting distribution of healthy and bankrupt firms in the training, validation and test sets can be seen in Fig. 2.

Document segmentation is carried out for long financial disclosures. After removing noise in the form of the HTML document, and common boilerplate text, each document gets tokenized and broken into fixed sized chunks. Let T be the number of tokens in a document and L be the chunk's max token length. The number of chunks N is calculated by dividing the document into segments of length L, as in Eq. (1). In our experiments the number of tokens in a chunk is 512, following the choice of the maximum sequence length of the transformer encoder. Each document is further reduced in size to a maximum of 20 chunks, to manage the memory

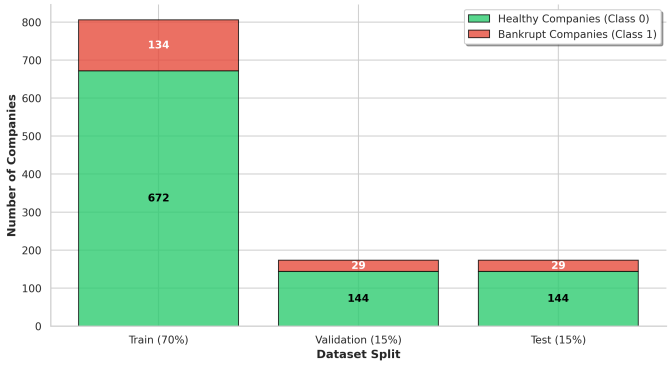


Fig. 2. Class distribution of healthy and bankrupt firms in the training, validation, and test sets.

usage while maintaining enough contextual information. While this memory-constrained limit truncates 67.01% of the longest documents and yields an overall truncation loss of 33.62%, it securely retains the most critical semantic cues and distress indicators, which typically appear in the earlier sections of management and auditor reports.

$$C = \min\left(C_{max}, \lceil \frac{T}{L} \rceil\right) \tag{1}$$

This chunked representation prevents hard truncation and supports hierarchical encoding.

C. Feature Engineering and Selection

A multivariate feature set is created to represent the accounting signals as well as the narrative risk cues. All columns other than UID and label, item 7, opinion text and the combination of raw text field are treated as candidate numerical predictors. Object type variables are label encoded, Infinite values are replaced by missing, Variables with missing rate higher than 0.60 are removed. Remaining missing entries are imputed using an iterative multivariate imputation maintaining cross feature dependencies.

To explicitly capture distress language, a unified disclosure field is formed by concatenating item 7 with audit opinion text. For a lexicon V and the token set W_i of document i, the density feature is defined in Eq. (2):

$$Density_{V,i} = \frac{\sum_{w \in W_i} I(w \in V)}{|W_i|} \tag{2}$$

Using negative, litigious, and uncertainty lexicons, we construct three density indicators, and a binary going concern flag is added when the phrase is present. These linguistic red flags strengthen interpretability and support later robustness reasoning.

Mutual information is then employed to generate the top 5 numerical predictors and subsequently only the interaction-only second-degree polynomials terms are produced. This expansion of the number of base predictors to 5 is the limit that is set so as to guarantee efficiency in the computation and to avoid overfitting. Any further extension would further raise the feature space exponentially, which is necessarily going to incur the curse of dimensionality and terrible multicollinearity, due to the extreme imbalance between classes in the case of bankruptcy data. The model only uses the top 5 predictors which is a sufficient approach to capturing the most statistically significant non-linear relationships with no unnecessary noise being modeled. For two selected predictors x_a and x_b the interaction is defined in Eq. (3):

$$x_{ab} = x_a x_b \tag{3}$$

All the numerical predictors are clipped at 1st and 99th percentile and transformed into a normal like distribution through quantile transformation. A final top K predictor is then selected by mutual information which is used by the model. The sensitivity analysis is done to find the best feature selection and test the strength of the model by changing the value of K (e.g., K = 10, 20, 30, and 40). The last model uses K=20 due to the maximum overall validation performance. The impact of this option is also addressed in the experimental results.

D. Model Architecture of the Hybrid D-HAN-Net

The proposed architecture is designed to learn stable numerical structure, preserve long range disclosure evidence, and align both modalities through attention guided fusion. The pipeline contains four stages.

1) Hierarchical numerical encoder: The GRN framework supports nonlinear transformation and performs the original feature representation by residual connections. Financial ratios tend to be heterogeneously distributed and non linearly interacting. Thus, their predictive structure cannot be well represented by a simple linear transformation. In order to stabilize the learning process, the encoder uses two nonlinear transformation layers and then dropout regularization and projection residual. A gated residual network is used to project the selected predictors to an embedding of 128 dimensions. Let r be the projected residual and $f(x)$ be a nonlinear transform, then the numerical representation is defined in Eq. (4). The hidden representation dimension is set to 128 in our experiments.

$$h_{num} = BN(r + g \odot f(x)) \tag{4}$$

where, $g = \sigma(W_g f(x))$ and $\odot$ denotes element-wise multiplication. This gating mechanism enables the network to adaptively suppress noisy financial indicators while amplifying features that strongly correlate with bankruptcy risk. Batch normalization further stabilizes training by normalizing the representation distribution.

2) Hierarchical text encoder: The FinBERT application is informed by the fact that corporate disclosure is a domain-structured linguistic. Accounting terminology, expressions of financial risk, and statements of uncertainty are common management discussion narrative language that is not well reflected in general language corpora. FinBERT is trained on financial text, and thus it picks up contextual semantics to do with financial distress, compared to general-purpose BERT models. The GRU layer is bidirectional and additional to this, the layer

performs segment level embedding on the entire document and this identifies long-range dependencies among disclosure segments. The contextual embedding of each segment is obtained from the pooled output of FinBERT, which is chosen over standard BERT models to accurately interpret the specialized financial terms and unique phrasing found in management narratives. These segment embeddings are then organized as a sequence and passed through a bidirectional GRU layer. The bidirectional recurrent layer aggregates contextual information from both earlier and later segments, enabling the model to capture narrative dependencies across the processed sequence.

A bidirectional GRU aggregates chunk embedding into a single 128-dimensional representation h_{text}. The FinBERT parameters are frozen during training to reduce computational cost and prevent overfitting on the relatively small bankruptcy dataset. The resulting sequence representation is aggregated by the GRU layer to form a document-level embedding. Compared with fully transformer-based long-document architectures, the bidirectional GRU offers lower computational complexity and memory overhead. It also provides stable sequence aggregation for small and highly imbalanced financial datasets.

3) Cross-modal fusion with gating: Once numerical and textual representations are obtained, the next stage aligns both modalities through a cross-modal attention mechanism. Bankruptcy signals often emerge when textual statements confirm or contradict financial ratios. The fusion module therefore enables interaction between these two sources. The cross-attention module uses the summarized textual representation h_{text} as a Query to attend to the numerical representation h_{num}, acting as Key and Value, producing an attention-refined text vector h_{attn}. A learnable fusion gate then adaptively blends both streams, as defined in Eq. (5):

$$h_{fuse} = z \odot h_{num} + (1 - z) \odot h_{attn} \tag{5}$$

where, $z = \sigma([h_{num}; h_{attn}])$, the gate value z lies in the range 0 to 1, where values closer to 0 indicate stronger reliance on textual signals, whereas values closer to 1 indicate greater reliance on numerical indicators. Here, σ denotes the sigmoid activation function, $[;]$ denotes concatenation, and $\odot$ indicates element wise multiplication. This mechanism allows the model to adaptively decide whether bankruptcy evidence should rely more strongly on financial ratios or on textual disclosure signals for each firm. To test the effectiveness of the proposed cross-modal fusion gate, an easier to implement late fusion baseline is also considered. In this alternative environment the numerical and textual representations are simply concatenated and fed to the classifier without interaction between the two representations based on attention. This comparison is helpful for determining if the proposed fusion mechanism offers a meaningful advantage over a simpler multimodal integration strategy.

4) Probabilistic prediction and decision logic: The fused representation is finally passed through a shallow classifier consisting of a fully connected layer with GELU activation followed by dropout regularization. This stage converts the fused feature vector into a scalar bankruptcy probability. The final prediction is obtained using a sigmoid activation, defined in Eq. (6):

$$p = \frac{1}{1 + e^{-(w^\top h_{fuse} + b)}} \tag{6}$$

where, p represents the predicted probability that the firm will experience bankruptcy in the next fiscal period. The sigmoid function ensures that the output remains within the range of zero to one. During evaluation, the predicted probability is compared with an optimal decision threshold determined from the validation precision recall curve. This adaptive threshold selection helps balance precision and recall under severe class imbalance.

E. Training Strategy and Optimization

The training procedure is designed to remain stable under extreme class imbalance and long document variability, while ensuring that improvements in AUC and F1 are not driven by a single favorable split. The overall objective is to learn a calibrated bankruptcy probability that is robust across random initializations and sampling noise.

1) Imbalance-aware training and loss optimization: Bankruptcy samples are rare, hence standard mini batch sampling would lead to overfitting of the majoritarians of the healthy sample class by the optimizer. To address this, a technique of weighted random sampling is applied, at the instance level. Each training example is assigned a weight inversely proportional to its class frequency and batches are drawn with a WeightedRandomSampler. This makes the effect of bankrupt instances more effective without duplicating them deterministically, so the gradient bias for the negative class decreases. Model parameters are optimized using AdamW, which combines adaptive learning rates with decoupled weight decay.

A learning rate of 1.5×10^{-5} is selected to maintain stable updates when training with frozen language representations and a relatively small classifier head. Dropout is applied in the numerical encoder, fusion feedforward layers, and classifier head to reduce co adaptation between modalities and to prevent the fusion gate from collapsing to a single modality. Even with balanced sampling, many bankrupt signals remain subtle, so the training objective must prioritize difficult cases. The model is trained using focal loss, which down weights easy samples and focuses learning on hard or misclassified examples. With $y \in \{0, 1\}$ and predicted probability p, define $p_t = p$ if $y = 1$ else $1 - p$. The loss is defined in Eq. (7):

$$L_{focal} = -\alpha(1 - p_t)^\gamma \log(p_t) \tag{7}$$

The focusing parameter is set to $\gamma = 2$ in our experiments. The balancing factor $\alpha = 0.25$ is used to assign higher importance to the minority bankrupt class.

2) Validation protocol and stability analysis: Training runs are for up to 40 epochs, but performance is not assumed to improve monotonically. The validation AUC is calculated after every epoch with the help of predicted probabilities. The model checkpoint with the highest validation AUC value is saved and early stopping is triggered when there is no improvement in validation AUC for 10 consecutive epochs. This helps avoid over training the classifier head and the fusion layers which

is particularly common when the text encoder is frozen and the capacity is focused on fusion block. Since the dataset is imbalanced, applying a default classification threshold of 0.5 often yields a high AUC but a poor F1-score for the positive class.

Therefore, after being trained, the decision threshold is selected by maximizing the F1 score over the precision recall curve, which results in an equilibrium operating point between false alarms and missed bankruptcies. This threshold is then applied to the held out test set for final reporting. In order to guarantee the robustness, all experiments are repeated under multiple random seeds, which alters the selections of healthy samples and initializations of fusion and classifier parameters. Final results are presented in terms of the mean and standard deviation, which indicate the stability of the model with regard to sampling uncertainty and randomness of optimization.

F. Evaluation Protocol and Robustness Reporting

The assessment plan will be used to facilitate fair comparison of the offered model and the baseline methods. Stratified sampling is used to partition the dataset into training, validation and test sets such that the classes are always distributed equally across the partitions. The selection of models is done on the basis of validation performance in which the primary monitoring measure is the Area Under the ROC Curve. Due to the high imbalance in bankruptcy prediction, it does not mean that the eventual classification point is 0.5. Rather, the maximum value of the F1 score on the validation set is chosen and used on the held out test set. This technique offers a better balance between trade off in precision and recall. In order to measure robustness, each experiment is run with a number of random seeds and the final value is presented as mean and standard deviation. Besides, the suggestion model is contrasted with the methods of strong baseline to demonstrate the efficiency of the multimodal architecture.

IV. Results and Discussion

A. Experimental Configuration and Implementation Details

All experiments have been implemented in Python. Numerical baselines were built using Scikit-learn, the proposed multimodal model was implemented using PyTorch, and transformer baselines were implemented using HuggingFace Transformers. Experiments were done on the Kaggle research platform with GPU/CUDA to speed up the training and inference. Data preprocessing and feature engineering were done using Pandas, NumPy and Scikit-learn. Performance was measured in terms of Accuracy, Precision, Recall, F1-score and AUC. To make the comparison as broad as possible, baseline models were developed for both the numerical and textual modalities. The number baselines were Random Forest 100 trees, XGBoost logistic objective, SVM probabilistic output, and a 2-layer multilayer perceptron with 64 and 32 hidden layers. The textual baselines were BERT-Base, RoBERTa, DistilBERT, and ELECTRA fine-tuned for three epochs and batch 16 maximum sequence length 128.

D-HAN-Net was optimized using AdamW with learning rate of 1.5×10^{-5}, batch size 16 and max. 40 epochs. Early stopping was used according to validation AUC using patience of 10 epochs. To deal with the class imbalance, a weighted random sampling was adopted in the mini-batch construction and focal loss was taken into consideration to give priority to difficult samples. The final classification threshold was chosen from the precision-recall curve by finding the F1-score (Table I).

TABLE I. Key Training Configuration for the Proposed D-HAN-Net Model.

Parameter	Value
Healthy:Bankrupt Ratio	5:1
Data Split (Train/Val/Test)	70% / 15% / 15%
Selected Numerical Features	20
Text Encoder Aggregation	Bidirectional GRU
Numerical Representation	Gated Residual Network (GRN)
Multimodal Fusion	Cross-modal Attention + Gate
Optimizer & Learning Rate	AdamW, 1.5×10^{-5}
Batch Size & Max Epochs	16, 40
Imbalance Handling	Weighted Sampling, Focal Loss ($\gamma = 2$)

B. Performance Evaluation and Empirical Analysis

Across all baselines, the test-set results confirm that bankruptcy signals are captured much more reliably by structured financial variables than by narrative text alone, as illustrated in Table II. Among numerical baselines, MLP achieves the best overall score with accuracy 0.9133 and AUC 0.9672 and SVM and Random Forest closely follow with accuracy close to 0.89-0.90 and AUC above 0.94. However, Table II also shows a clear trade-off in which XGBoost and SVM have the highest recall of 0.9310 but their precision remains lower with 0.5510 and 0.6136, which implies a higher number of false positive bankruptcy alarms. The transformers that are text-only perform a weaker job overall, with accuracy maximum of 0.8324 and AUC mostly in between 0.61-0.69 which means that separability with text only is limited. D-HAN-Net is unique in the way that it enhances the overall ranking and classification simultaneously. It has the highest accuracy of 0.9400 and the best AUC of 0.9709, which shows the most obvious separation between bankrupt and healthy firms across thresholds in Table II. At the same time, it maintains precision of .7485 and recall of .8828 which means the model does not face the common tradeoff between high recall and many false positives. This balance can be seen in its top F1-score in comparison with all baselines. The result is consistent with recent multimodal bankruptcy studies which indicate that combining financial and textual information improves minority distress detection [36], [37].

Table III provides a comparison between the AUC scores of the model with different sets of features (K). When only 10 features were used, the AUC is less (0.8997) since the model has limited depth with regards to finances. Maximum performance is achieved at the top of the optimal K = 20 (our proposed model) which captures the most effective and relevant financial signs of risk assessment. Further increases in AUC to 30 and 40 bring the values to 0.9373 and 0.9303, respectively. This deterioration confirms that too many variables bring about noise and redundancy of data which ultimately diminishes the total accuracy of prediction.

Fig. 3 reports the confusion matrix of D-HAN-Net on the held out test set and explains how the final accuracy

TABLE II. BASELINE COMPARISON RESULTS ON THE TEST SET

Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
Numerical	Random Forest	89.60%	64.10	86.21	73.53	95.31
Numerical	XGBoost	86.13%	55.10	93.10	69.23	93.92
Numerical	SVM	89.02%	61.36	93.10	73.97	95.47
Numerical	MLP	91.33%	68.42	89.66	77.61	96.72
Text	BERT-Base	83.24%	50.00	37.93	43.14	68.51
Text	DistilBERT	67.63%	27.87	58.62	37.78	66.55
Text	ELECTRA	73.99%	30.95	44.83	36.62	61.09
Numerical +Text	D-HAN-Net	94.00%	86.00	92.00	88.00	97.34

TABLE III. AUC PERFORMANCE BASED ON FEATURE SELECTION (TOP_K).

Number of Selected Features (TOP_K)	AUC Score
TOP_K = 10	0.8997
TOP_K = 20	0.9734
TOP_K = 30	0.9373
TOP_K = 40	0.9303

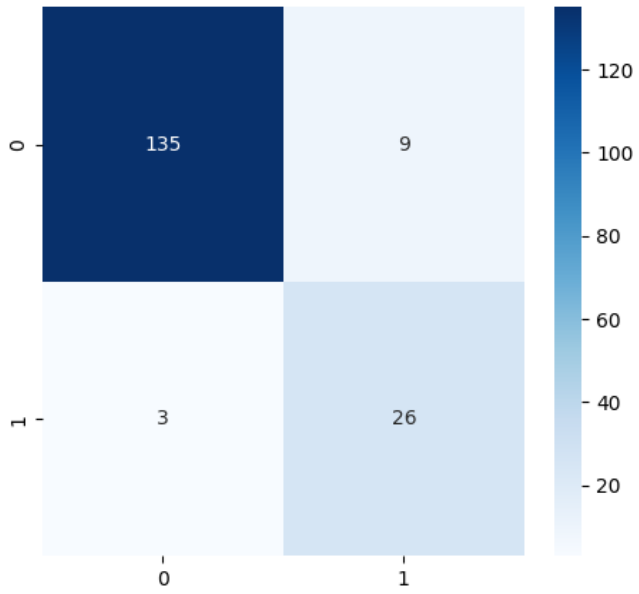


Fig. 3. Confusion matrix of the D-HAN-Net on the test dataset.

gets formed using class wise outcomes. The confusion matrix shows that 135 healthy firms are correctly classified as healthy and 26 bankrupt firms are correctly identified as bankrupt. The remaining errors are limited, and asymmetric in a good way for use in early warning. Only 9 healthy firms are flagged as bankrupt which represents false alarms which may trigger unnecessary investigation, while only 3 bankrupt firms are classified as healthy which represents missed distress cases which are usually more costly in practice. The low number of false negatives shows that the model captures most of the distressed firms, and does not depend on the majority class to inflate the accuracy. At the same time, a low number of false positives means the model does not create over sensitive screening that will overwhelm stakeholders with alerts. Overall, in the matrix, D-HAN-Net is seen to have a practical balance between caution and precision that generates high true positive detection and keeps the misclassification rates under control. This behavior is consistent with a robust multimodal decision process that relies on financial ratios, as well as disclosure signals.

C. Threshold Independent Performance Evaluation

Instead of improving performance at a single fixed threshold, D-HAN-Net provides a predictable better performance balance across a wide range of operating points, and this can be seen from both of the curves in Fig. 4 and Fig. 5. In Fig. 4, its ROC trajectory remains most close to the top left region over most of the false positive range, which represents the best separating capability between bankrupt and non-bankrupt firms. The early rise of the curve indicates that the sensitivity is gained rapidly and the small drift to the right indicates that false positives are still under control. The numerical baselines lag behind them with very high but less high curvature so their extra true positive gain tends to demand more false positives. The text-only transformers are still closer to the diagonal trend, with very limited discriminative power across thresholds in Fig. 4. This suggests that narrative disclosures alone may capture uncertainty-related semantic cues but remain insufficient for stable bankruptcy discrimination without supporting financial indicators. D-HAN-Net is unique in that it enhances overall ranking as well as classification together. It has the highest accuracy of 0.9400 and the best AUC of 0.9734 demonstrating the clearest separation of bankrupt and healthy firms for different thresholds, as in Table II. At the same time, it maintains the precision level at 0.86 and recall level at 0.88 and thus the model does not suffer from the usual tradeoff where higher recall comes with a lot of false positives. This balance is reflected in its top F1-score compared with all the baselines. The improved performance likely arises from the complementary interaction between structured financial indicators and contextual narrative evidence under highly imbalanced distress conditions.

The same advantage passes on to the more decision critical view in Fig. 5. D-HAN-Net maintains precision when recall is increased, thus can recall more bankrupt firms without quickly raising false alarms. This is important in screening tasks where it is costly to miss a bankrupt case, but too many false alerts will also make the tool less practical. Several numerical baselines exhibit a steeper drop in precision at higher recall, indicating that the sensitivity (i.e. pushing to create more alarms) produces more noisy alarms. The text-only models fall to low precision early, and are weak throughout the curve, indicating that the narrative signals alone are insufficient to

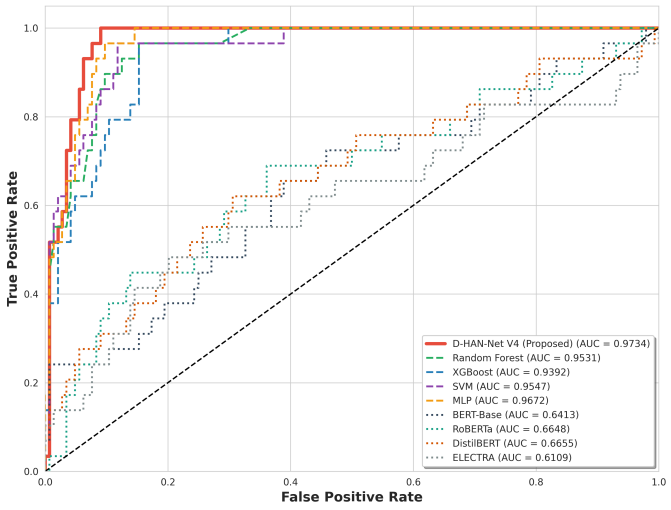


Fig. 4. Receiver Operating Characteristic (ROC) curves for all models.

secure stable retrieval.

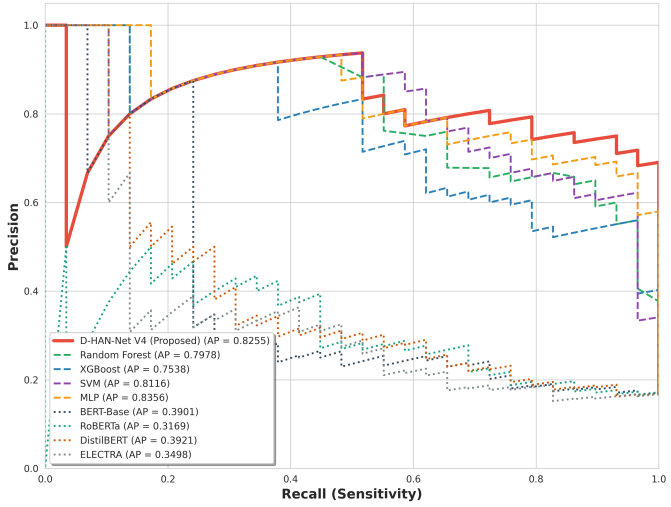


Fig. 5. Precision–Recall (PR) curves for all models.

Taken together, Fig. 4 and Fig. 5 imply that D-HAN-Net has strong ranking quality and stable class retrieval, simultaneously, and the advantage is retained in the whole threshold space instead of only one chosen decision point, which is helpful for deployment under various risk preferences.

D. Model Robustness and Cross Seed Stability Analysis

Robustness analysis evaluates whether D-HAN-Net stays reliable when training is repeated with different random seeds and compares this stability against strong numerical baselines in Table IV. The goal is to ensure that the reported gains are not caused by one favorable run, but remain consistent across multiple executions of the pipeline. Table IV indicates that D-HAN-Net achieves the best overall performance across five seeds. It records the highest accuracy at 94.00 $\pm$ 0.57 and the strongest AUC at 97.34 $\pm$ 0.35, showing superior separation between bankrupt and healthy firms. The F1 score is also the highest at 88.00 $\pm$ 0.75, supported by high precision at 86.00 $\pm$ 4.14 and high recall at 92.00 $\pm$ 4.68, which reflects balanced detection rather than a one sided gain. The error profile further strengthens this conclusion in Table IV. D-HAN-Net shows the lowest Type I error at 6.11 $\pm$ 1.55, which means fewer false bankruptcy alarms than XGBoost, Random Forest, and SVM. Type II error remains controlled at 11.72 $\pm$ 4.68, while Random Forest shows a higher Type II error at 17.93 $\pm$ 4.02, implying more missed bankrupt cases. XGBoost and SVM reach very high recall at 93.10, but their lower precision and higher Type I error indicate a larger false positive burden. Overall, Table IV supports that combining auditor text with financial ratios produces a more stable and dependable predictor than single modality baselines.

E. Component Contribution and Ablation Investigation

The ablation study is to break the full D-HAN-Net into key parts, to quantify the contribution of each block to the prediction quality. This type of analysis helps validate the components, as well as measuring the synergy, that the overall gains are the result of the combined design, rather than one block alone. As presented in Table V, the result of the full model is the best, which is accuracy 94.00, F1-score is 88.00 and AUC is 97.34. When GRN is removed, performance deteriorates drastically to accuracy 83.82 and AUC 84.77, suggesting that gated refinement is central for strong representation learning. Removing sequential aggregation also decreases performance, with accuracy 84.97 and F1 64.79, which suggests that sequence level cues help the model form a more stable decision. When attention fusion is removed, F1 drops to 61.76, AUC drops to 86.69, this means that cross-modal alignment is required to maintain a balance between the predictions. The biggest degradation occurs when focal loss is removed, in which F1 will be 60.38 and AUC will be 83.33, indicating the importance of the loss design in handling class imbalance and avoiding unstable learning. Overall, Table V suggests that strong and consistent results demand the complete integration of GRN, sequential modeling, attention fusion and focal loss, as opposed to some simplified variant.

F. Model Interpretability Through Feature Impact Analysis

A clear hierarchy of risk drivers can be seen from the SHAP summary in Fig. 6. Going Concern Opinion is the best contributor, in which higher feature values drive the predictions into the bankrupt category and lower values are more likely to be healthy companies. The extensive dispersion of SHAP values indicates that audit opinion is a powerful separating signal particularly to firms whose explicit distress is enduring. The second and third in pecking order are retained earnings and variables that are associated with net income, which implies that reduced profitability and exhausted accumulated earnings add to the risk of bankruptcy. Most of their SHAP effects are skewed towards the positive in unfavorable values indicating that the model considers the weaknesses of earnings as a direct distress signal. It is also indicated by Fig. 6 that the model is based on the mix of audit signals and core indicators of financial strength instead of the particular ratio. Balance sheet variables like current assets and total stockholders equity have a similar direction with less financial

TABLE IV. ROBUSTNESS ANALYSIS ACROSS 5 RANDOM SEEDS

Model	Accuracy	AUC	F1-Score	Precision	Recall	Type I Error
XGBoost	86.13 ± 0.00	93.92 ± 0.00	70.89 ± 0.00	55.10 ± 0.00	93.10 ± 0.00	15.28 ± 0.00
Random Forest	90.06 ± 0.77	94.95 ± 0.36	74.93 ± 0.75	66.78 ± 3.33	82.07 ± 4.02	8.33 ± 1.58
SVM (RBF)	89.02 ± 0.00	95.57 ± 0.11	75.68 ± 0.00	61.36 ± 0.00	93.10 ± 0.00	11.81 ± 0.00
D-HAN-Net (Proposed)	94.00 ± 0.57	97.34 ± 0.35	88.00 ± 0.75	86.00 ± 4.14	92.00 ± 4.68	6.11 ± 1.55

TABLE V. ABLATION STUDY OF D-HAN-NET: IMPACT OF KEY COMPONENTS ON PERFORMANCE.

Model Variant	Accuracy	F1-Score	AUC
D-HAN-Net	94.00%	88.00	97.34
w/o GRN	83.82%	61.97	84.77
w/o Sequential Agg (Bi-GRU)	84.97%	64.79	89.22
w/o Attention Fusion	84.39%	61.76	86.69
w/o Focal Loss	87.28%	60.38	83.33

strength being associated with increased risk. Other accounting indicators such as federal tax reconciliation and comprehensive income also have effects on prediction, but the average effects are smaller. This indicates that the model reflects primary distress indicators and secondary accounting indicators linked to declining performance. Altogether, the feature effects are economically sound. The model relates the indicators of audit-based concerns with the profitability and solvency indicators within one interpretable decision framework, with the highest impact on the audit opinion, then the quality of earnings, and the support provided by the resilience of the balance sheet, respectively, to the ultimate prediction.

G. Case-Specific Evaluation and Linguistic Interpretability

This case study highlights how confidently each of the models flags a difficult bankrupt firm in Table VI. D-HAN-Net has the highest bankrupt confidence of 61.67, showing the clearest risk judgment of this sample among all methods. SVM is close at 60.93, while XGBoost is lower at 55.18, which seems to exhibit lower level of commitment despite risk detection. A larger gap is seen for the rest of the baselines in Table VI. Random Forest drops to 44.00 and MLP drops even further to 32.07, meaning that they consider the same firm less decisively bankrupt. The lowest scores are experienced by the text only transformers, hovering around 13.12 to 16.38, suggesting that they mostly fail to pick up on the evidence of distress in this difficult example. Overall, Table VI presents that D-HAN-Net is the most decisive model on a hard case, where single modality baselines tend to under estimate the bankruptcy risk.

Fig. 7 illustrates the most common linguistic cues to financially struggling firms taken from Management Discussion and Analysis and auditor opinion sections. The word cloud emphasises such dominant words as costs, credit, interest, sales, capital, and accounts that are often found in bankruptcy related disclosures. The prominence of these terms reflect common financial distress narratives including rising operational costs, debt obligations, declining revenues, and restructuring pressures. This visualization is qualitative evidence that there are meaningful warning signals of corporate failure in the textual disclosures. Such linguistic patterns complement the

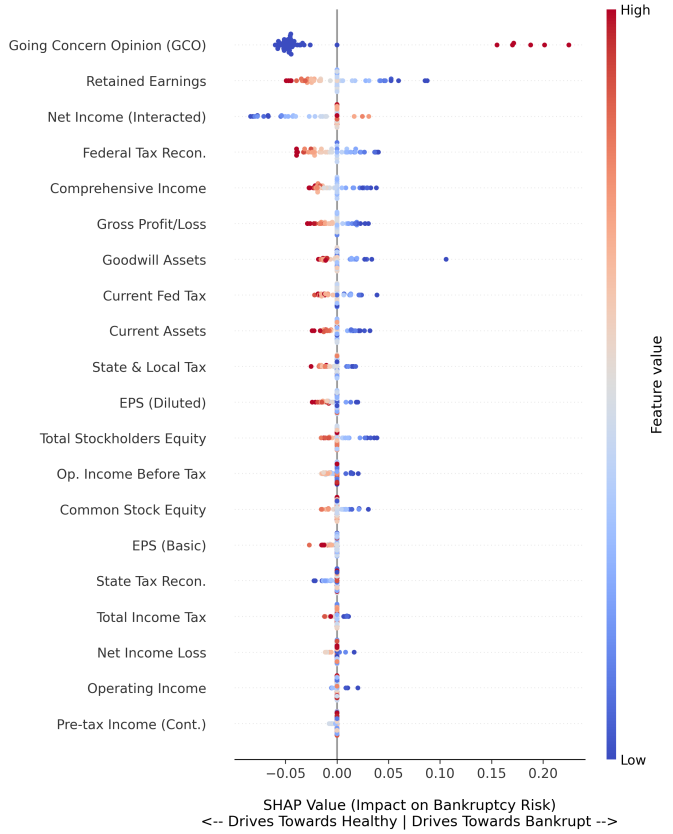


Fig. 6. SHAP Summary plot: Evaluating global feature importance and directional impact in D-HAN-Net.

TABLE VI. CRITICAL CASE STUDY: CROSS-MODEL BANKRUPTCY CONFIDENCE SCORES (TEST INDEX = 45).

Model Architecture	Confidence Score (Bankrupt)
SVM	60.93
XGBoost	55.18
Random Forest	44.00
MLP	32.07
BERT-Base	15.49
ELECTRA	16.38
DistilBERT	13.12
D-HAN-Net (Proposed)	61.67

numerical financial indicators that are to be used in the proposed framework and help to explain the role of textual evidence in predicting bankruptcy. The analysis supports the interpretability objective it shows that the model captures real world distress related language present in corporate reports rather than relying solely on statistical correlations.

TABLE VII. BENCHMARK COMPARISON OF EXISTING MODELS WITH THE PROPOSED D-HAN-NET

First Author	Year	Architecture	Dataset	AUC
Ainan et al. [34]	2024	Hybrid Machine Learning	Unbalanced Polish Dataset	0.9580
Mancisidor et al. [37]	2022	CMMD	U.S. Firms	0.9327
Fasano et al. [32]	2024	Explainable AI	Corporate Financial Dataset	0.9270
Arno et al. [36]	2023	Multimodal Ensemble	ECL Dataset	0.9150
Correia [35]	2025	Traditional Accounting Models	CRSP U.S. Firms	0.9136
Xiao et al. [33]	2024	DGCN-TA	Enterprise Relationship Networks	0.8867
Mai et al. [7]	2019	Deep Learning	U.S. Public Companies	0.8560
D-HAN-Net (Proposed)	2026	D-HAN-Net	Multimodal	0.9734



Fig. 7. Linguistic red flags extracted from bankrupt firms' auditor reports.

Table VII indicates that the proposed D-HAN-Net is an effective tool in comparison to some other recent studies on bankruptcy prediction. Prior deep learning studies like Mai et al. had already achieved an AUC of 0.8560, which points to a moderate level of predictive performance in the case of using textual disclosures as the main predictive factor. Subsequent methods were able to enhance performance by adding more structural or financial data. As an illustration, the DGCN-TA model by Xiao et al. resulted in an AUC of 0.8867 with the classical accounting based model by Correia et al. having an AUC of 0.9136. Further developed architectures also improved the performance of prediction. Arno et al. multimodal ensemble method achieved an AUC of 0.9150, and Fasano et al. showed the applicability of explainable AI of 0.9270. On the same note, the CMMD model of Mancisidor et al. scored 0.9327, and the hybrid machine learning model of Ainan et al. scored 0.9580, which is among the highest scores ever seen among previous researchers. Although these gains have been made, the proposed D-HAN-Net has the best AUC of 0.9734, which is obviously above all current methods, as summarized in Table VII. This outcome shows that the hierarchical multimodal structure better extracts intricate bankruptcy indicators from textual and financial features. By combining these complementary modalities, the proposed framework achieves superior discrimination and more reliable predictive performance. This strength is evidenced by the model's robustness and consistently strong predictive capability across numerous experiments.

V. CONCLUSION

A robust early warning system for corporate financial distress is essential for investors, regulators, and financial institutions to manage risk and preserve market stability. This study developed D-HAN-Net, a hybrid dual-stream deep learning architecture that integrates structured financial indicators with textual signals derived from corporate disclosures. By jointly modeling numerical financial dynamics and narrative evidence from MD&A sections and auditor reports, the proposed framework captures complementary distress signals that are often overlooked in single-modality approaches.

The empirical results show that the combination of financial ratios and textual information from written disclosure is able to improve the bankruptcy prediction performance significantly. The proposed model has the highest performance in terms of all evaluation metrics such as accuracy, AUC, and F1-score with an efficient balance between precision and recall. Compared with the standard machine learning benchmarks and the transformer-based text models, D-HAN-Net is able to reduce classification errors and increase the overall classification reliability. Moreover, the fact that results are consistent across multiple random seeds suggests that the observed performance gains are robust, rather than sample-specific. The ablation analysis further indicates that every architectural component has a material contribution to the overall predictive performance of the model, with especially high contribution of the gated residual numerical encoder, hierarchical text representation, and cross-modal fusion mechanism. These results highlight the importance of bringing diverse financial information together in a cohesive predictive model. In addition, the interpretability analysis indicates that the model captures economically meaningful indicators of distress. Audit-related going concern indicators as well as profitability and balance sheet strength indexes become key variables. The linguistic analysis of disclosure texts also suggests that the model finds distress-related narrative patterns often found with financially vulnerable firms.

Beyond the predictive performance, the results of this study have meaningful implications for the financial reporting regulation. Narrative disclosures, especially MD&A discussions and auditor reports, provide informative signals of financial distress. Therefore, regulatory authorities may wish to consider enhancing mandatory reporting standards to increase the clarity, consistency and decision-usefulness of such disclosures. More standardized and informative reporting practices could help improve the effectiveness of data-driven early warning systems and provide support to more timely oversight by regulators, creditors and investors.

Overall, the combination of structured financial data and narrative disclosure is a more complete representation of cor-

porate financial condition. The proposed architecture provides a workable and reliable framework for early prediction of bankruptcy, as well as informed financial risk assessment. Despite the strong predictive performance, the proposed framework may still be affected by temporal distribution shifts, industry-specific reporting variations, and incomplete disclosure availability in real-world financial environments. Future research may investigate temporal forecasting settings and robustness under changing economic conditions.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Southeast University, Dhaka, Bangladesh, for providing the essential laboratory facilities and technical support required to conduct this research.

REFERENCES

[1] Y. Ji, X. Zhou, Z. Cai, and Z. Yin, "Enhancing status-based optimization via dual dispersal mechanisms for feature selection in bankruptcy prediction," *Neurocomputing*, vol. 668, p. 132394, 2026. doi: 10.1016/j.neucom.2025.132394.

[2] Y.-C. Lin, R. Padliansyah, Y.-H. Lu, and W.-R. Liu, "Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques," *International Journal of Accounting Information Systems*, vol. 56, p. 100744, 2025. doi: 10.1016/j.accinf.2025.100744.

[3] W. Liu, H. Wu, T. Wang, et al., "Elite elimination osprey optimization algorithm optimized kernel extreme learning machine for bankruptcy prediction problems," *Scientific Reports*, vol. 16, p. 6246, 2026. doi: 10.1038/s41598-026-37249-9.

[4] K. Jiang, X. Zhao, Y. Li, et al., "A thinking innovation strategy based Northern goshawk optimizer enhanced extreme learning machine for bankruptcy prediction problems," *Scientific Reports*, vol. 16, p. 9628, 2026. doi: 10.1038/s41598-025-34452-y.

[5] G. Lombardo, M. Pellegrino, G. Adosoglou, S. Cagnoni, P. M. Pardalos, and A. Poggi, "Machine learning for bankruptcy prediction in the American stock market: Dataset and benchmarks," *Future Internet*, vol. 14, no. 8, p. 244, 2022. doi: 10.3390/fi14080244.

[6] G. Ranjbaran, D. R. Recupero, G. Lombardo, and S. Consoli, "Leveraging augmentation techniques for tasks with unbalancedness within the financial domain: a two-level ensemble approach," *EPJ Data Science*, vol. 12, no. 1, p. 24, 2023. doi: 10.1140/epjds/s13688-023-00402-9.

[7] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *European Journal of Operational Research*, vol. 274, no. 2, pp. 743–758, 2019. doi: 10.1016/j.ejor.2018.10.024.

[8] W. J. Mayew, M. Sethuraman, and M. Venkatachalam, "MD&A disclosure and the firm's ability to continue as a going concern," *The Accounting Review*, vol. 90, no. 4, pp. 1621–1651, 2015. doi: 10.2308/accr-50983.

[9] F. Li, "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and Economics*, vol. 45, no. 2-3, pp. 221–247, 2008. doi: 10.1016/j.jacceco.2008.02.003.

[10] T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[12] T.-K. Chen, H.-H. Liao, G.-D. Chen, W.-H. Kang, and Y.-C. Lin, "Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports," *Expert Systems with Applications*, vol. 233, p. 120714, 2023. doi: 10.1016/j.eswa.2023.120714.

[13] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, p. 589, 1968. doi: 10.2307/2978933.

[14] S. P. Kothari, X. Li, and J. E. Short, "The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis," *The Accounting Review*, vol. 84, no. 5, pp. 1639–1670, 2009. doi: 10.2308/accr.2009.84.5.1639.

[15] F. Li, "The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach," *Journal of Accounting Research*, vol. 48, no. 5, pp. 1049–1102, 2010. doi: 10.1111/j.1475-679X.2010.00382.x.

[16] H. Kim, H. Cho, and D. Ryu, "Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data," *Computational Economics*, vol. 59, no. 3, pp. 1231–1249, 2022. doi: 10.1007/s10614-021-10126-5.

[17] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert Systems with Applications*, vol. 117, pp. 287–299, 2019. doi: 10.1016/j.eswa.2018.09.039.

[18] M. Pellegrino, G. Lombardo, G. Adosoglou, S. Cagnoni, P. M. Pardalos, and A. Poggi, "A multi-head LSTM architecture for bankruptcy prediction with time series accounting data," *Future Internet*, vol. 16, no. 3, p. 79, 2024. doi: 10.3390/fi16030079.

[19] A. B. Hussain, S. Jingchun, B. Hussain, et al., "Cross-market bankruptcy prediction: An interpretable ensemble learning framework using SHAP analysis," *Computational Economics*, 2026. doi: 10.1007/s10614-025-11266-8.

[20] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *European Journal of Operational Research*, vol. 252, no. 2, pp. 561–572, 2016. doi: 10.1016/j.ejor.2016.01.012.

[21] D.-N. Wang, L. Li, and D. Zhao, "Corporate finance risk prediction based on LightGBM," *Information Sciences*, vol. 602, pp. 259–268, 2022. doi: 10.1016/j.ins.2022.04.058.

[22] S. Carta, A. Ferreira, D. R. Recupero, M. Saia, and R. Saia, "A combined entropy-based approach for a proactive credit scoring," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103292, 2020. doi: 10.1016/j.engappai.2019.103292.

[23] S. Carta, G. Fenu, A. Ferreira, D. R. Recupero, and R. Saia, "A two-step feature space transforming method to improve credit scoring performance," in *A Two-Step Feature Space Transforming Method to Improve Credit Scoring Performance*, Springer, 2020, pp. 134–157. doi: 10.1007/978-3-030-66196-0_7.

[24] D. Liang, C.-F. Tsai, H.-Y. Lu, and L.-S. Chang, "Combining corporate governance indicators with stacking ensembles for financial distress prediction," *Journal of Business Research*, vol. 120, pp. 137–146, 2020. doi: 10.1016/j.jbusres.2020.07.052.

[25] W.-C. Lin, Y.-H. Lu, and C.-F. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Systems*, vol. 36, no. 1, 2019. doi: 10.1111/exsy.12335.

[26] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques," *Procedia Computer Science*, vol. 162, pp. 895–899, 2019. doi: 10.1016/j.procs.2019.12.065.

[27] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine learning for financial risk management: A survey," *IEEE Access*, vol. 8, pp. 203203–203223, 2020. doi: 10.1109/ACCESS.2020.3036322.

[28] G. Lombardo, G. Trimigno, M. Pellegrino, and S. Cagnoni, "Language models fine-tuning for automatic format reconstruction of SEC financial filings," *IEEE Access*, vol. 12, pp. 31249–31261, 2024. doi: 10.1109/ACCESS.2024.3370444.

[29] M. Papík and L. Papíková, "The possibilities of using AutoML in bankruptcy prediction: Case of Slovakia," *Technological Forecasting and Social Change*, vol. 215, p. 124098, 2025.

[30] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning*, 1997, pp. 179–186.

[31] A. Sideras, K. Bougiatiotis, E. Zavitsanos, G. Paliouras, and G. Vouros, "A multimodal dataset of financial disclosures, MD&A, and audit opinions with next-year bankruptcy labels," *figshare*, Dataset, 2025. doi: 10.6084/m9.figshare.30305341.v1.

[32] F. Fasano, C. Adornetto, I. Zahid, M. La Rocca, L. Montaleone, G. Greco, and A. Cariola, "The dilemma of accuracy in bankruptcy prediction: a new approach using explainable AI techniques to predict corporate crises," *European Journal of Innovation Management*, vol. 28, no. 11, pp. 1–22, 2025. doi: 10.1108/EJIM-06-2024-0633.

[33] Y. Xiao and X. Liu, "Dynamic graph convolutional network with temporal aggregation for corporate bankruptcy prediction," *IEEE Access*, vol. 14, pp. 14985–14995, 2026. doi: 10.1109/ACCESS.2026.3656673.

[34] U. H. Ainan, L. Y. Por, Y.-L. Chen, J. Yang, and C. S. Ku, "Advancing bankruptcy forecasting with hybrid machine learning techniques: Insights from an unbalanced Polish dataset," *IEEE Access*, vol. 12, pp. 9369–9381, 2024. doi: 10.1109/ACCESS.2024.3354173.

[35] M. Correia, "Accounting and corporate failure: the evolving role of accounting information in bankruptcy prediction," *Accounting and Business Research*, vol. 55, no. 5, pp. 510–537, 2025. doi: 10.1080/00014788.2025.2516273.

[36] H. Arno, K. Mulier, J. Baeck, and T. Demeester, "From numbers to words: Multi-modal bankruptcy prediction using the ECL dataset," in *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, 2023, pp. 11–21. doi: 10.18653/v1/2023.finnlp-2.2.

[37] R. A. Mancisidor and K. Aas, "Multimodal generative models for bankruptcy prediction using textual data," *arXiv preprint arXiv:2211.08405*, 2022.

A Reliability-Aware Visual-Inertial Odometry for Dynamic and Low-Texture Environments

Yelu Liu[1], Ruokun Qu[2]*, Mengcheng Xu[3], Chenglong Li[4], Hui Jiang[5]
Engineering Technology Training Center, Civil Aviation Flight University of China, Chengdu 610000, China[1]
College of Air Traffic Management, Civil Aviation Flight University of China, Chengdu 610000, China[2,3]
School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China[2,5]
College of Flight Technology, Civil Aviation Flight University of China, Guanghan 618307, China[4]

***Abstract*—Visual-inertial odometry (VIO) tends to degrade in aggressive dynamic and low-texture environments, where rapid motion, weak visual structure, and moving objects reduce the reliability of visual observations. This study presents TRAIL-VIO, a temporal reliability-aware visual-inertial odometry framework with line feature enhancement. The method estimates temporal observation reliability by combining semantic priors with IMU-based motion consistency, which allows a continuous and time-varying assessment of observation quality instead of frame-wise decisions. A reliability-aware point-line association scheme is also introduced, where inertial prediction is used to constrain feature matching and partially corrupted line segments are selectively retained. In addition, a reliability-guided marginalization strategy is applied to reduce the influence of unreliable visual constraints before they are incorporated into the prior. Experiments on the EuRoC MAV benchmark and a self-collected UAV dataset show that TRAIL-VIO achieves average RMSE values of 0.042 m and 9.51 m, respectively, outperforming representative baseline methods in dynamic and low-texture scenarios. Additional ablation, parameter-sensitivity, and runtime analysis further verify the contribution of the main modules, the robustness of the selected parameters, and the computational feasibility of the proposed framework.**

***Keywords*—Visual-inertial odometry; point-line feature fusion; dynamic environments; observation reliability estimation; marginalization; UAV navigation**

I. Introduction

Visual-inertial odometry (VIO) is widely used for autonomous navigation of micro aerial vehicles (UAVs), mobile robots, and other platforms in GPS-denied environments. By combining geometric constraints from visual measurements with short-term motion information from inertial sensors, VIO enables accurate and continuous state estimation [1], [2].

However, in real-world environments, VIO systems often operate under non-ideal conditions. Aerial platforms often undergo rapid motion, which easily introduces motion blur and large viewpoint changes. At the same time, many environments are weakly textured and contain moving objects. These factors not only degrade feature quality, but also make visual observations less reliable. As a result, incorrect data association may occur, leading to reduced estimation accuracy or even tracking failure. Improving robustness under such aggressive and dynamic conditions is still challenging.

Existing methods for improving VIO robustness mainly follow two directions. One line of work focuses on enhancing the estimator, typically through tightly coupled visual-inertial optimization. Another line introduces additional geometric features, such as line features, to complement point features [3]. These strategies do improve performance in difficult scenarios. However, they still depend heavily on the quality of front-end observations. When motion becomes aggressive, both point and line features can degrade before the back-end has a chance to compensate.

To address the limitations of the static-world assumption, recent studies have explored dynamic-scene SLAM and VIO methods based on semantic segmentation and geometric consistency checks [4], [5], [6]. These approaches aim to detect and filter out dynamic observations. Semantic cues help identify potentially movable objects, while geometric verification and IMU-based reasoning provide additional constraints. Together, they improve robustness to a certain extent.

While these dynamic-scene pipelines can remove obvious outliers, they still rely heavily on frame-wise filtering. This makes them sensitive to segmentation errors and short-term changes in object motion. The limitation becomes more evident in structurally rich environments. In low-texture scenes, line features often play an important role. Even partial occlusion or slight dynamic interference can cause an entire line segment to be discarded. As a result, useful structural information is unnecessarily lost [6]. Most existing methods handle uncertain observations only at the front end, typically through simple down-weighting. Their influence on the back-end is rarely considered. Without a temporal evaluation of observation reliability, these imperfect measurements can still enter the optimization process. Over time, they are absorbed into the marginalized prior, which gradually degrades the consistency of the sliding-window estimation [7].

These limitations suggest that the key issue is not only identifying dynamic observations, but also evaluating their reliability over time. In aggressive scenarios, semantic predictions, geometric residuals, and image quality can vary significantly across frames. Relying on frame-wise decisions is therefore insufficient. A more reasonable approach is to estimate observation reliability in a temporal and recursive manner, and incorporate it consistently into both front-end processing and back-end optimization.

To address these issues, we present TRAIL-VIO, a

*Corresponding author

reliability-aware framework tailored for aggressive dynamic environments. Instead of treating observation quality as a simple filtering problem at the front end, our method estimates it over time and uses it throughout the system. The estimated reliability is used in feature association, local structure handling, and marginalization. In this way, unreliable observations can be handled earlier, which helps maintain stable tracking while preserving long-term consistency. The main contributions of this study are summarized as follows:

- We introduce a recursive method to estimate temporal observation reliability. By combining semantic cues with IMU-based motion consistency, reliability is modeled as a time-evolving state rather than independent frame-wise decisions. This leads to more stable behavior under dynamic disturbances.
- We propose a reliability-aware front-end for point-line association. IMU prediction is used to guide matching, and partially corrupted line features are handled more carefully. Instead of discarding them completely, the method preserves valid segments when possible, which improves robustness in low-texture scenes.
- We develop a marginalization strategy that takes observation reliability into account. Observations are screened before being incorporated into the prior, reducing the influence of degraded measurements. This helps limit error accumulation and improves long-term estimation consistency.

The rest of this study is structured as follows: Section II reviews the related work. Section III outlines the system architecture. Section IV gives the details of our proposal. Section V details the experimental results, and Section VI concludes the study and discusses potential directions for future work.

II. Related Work

A. Visual-Inertial Estimation and Feature Representations

Visual-inertial odometry typically advances on two fronts, the underlying state estimator and the choice of visual features. Point-based systems remain the standard. Frameworks like VINS-Mono [1] and ORB-SLAM3 [2] provide reliable starting points for most state estimation tasks. Yet, points alone often struggle in environments lacking distinct textures. To address this, systems such as PL-VIO [3] bring in lines or other geometric primitives, showing noticeable gains in structurally challenging scenes.

Early VIO architectures relied heavily on filtering techniques. MSCKF [8] set a benchmark here by introducing a multi-state constraint Kalman filter for vision-aided navigation. Later variants like S-MSCKF [9] pushed this concept further to better handle aggressive motion. Alternatively, ROVIO [10] tied direct image alignment into an EKF framework, which helps when standard feature tracking fails. These filter-based methods are fast. They fit well on resource-constrained hardware. Still, their accuracy tends to drop if visual observations degrade even for a short period.

Over time, the field shifted heavily toward nonlinear optimization. Keyframe-based, tightly coupled optimization proved highly effective in systems like OKVIS [11]. More recently, frameworks like OpenVINS [12] and VINS-Fusion [13] have generalized these concepts, making sliding-window optimization a common setup for multi-sensor fusion. In mostly static or visually clear environments, these methods are highly accurate. The catch lies in the front end. They generally trust that incoming visual constraints are reliable. When a platform moves aggressively through degraded conditions, this assumption often breaks down.

Beyond tweaking the backend estimator, researchers look to richer feature representations to bridge these reliability gaps. Point-line fusions are particularly common. PL-SLAM [14] and PL-VINS [15] early on showed how lines naturally complement points where textures are sparse. This sparked a wave of extensions. Works like Trifo-VIO [16], PLI-VINS [17], LRPL-VIO [18], PLE-SLAM [19] and PL-CVIO [20] adapt point-line constraints to various sensor setups and scenarios. Others look even higher up the geometric hierarchy. StructVIO [21] leverages structural regularities common in man-made spaces, while Pop-up SLAM [22] incorporates planar structures. Ultimately, adding these higher-level primitives helps stabilize the system when simple point features fail.

While incorporating more geometric features certainly improves observability, it does not automatically guarantee reliability. Real-world conditions, such as sudden lighting shifts, partial occlusions, or motion blur, can easily corrupt both point and line tracking. Lines present a unique challenge here. Most existing frameworks treat them as rigid, all-or-nothing entities. A line segment is typically either fully accepted into the optimization or discarded entirely, even if only a small section is corrupted. This binary approach often wastes valuable geometric data in cluttered environments.

Simply having more constraints available is not enough. A robust system must also be able to gauge which constraints to actually trust over time. This becomes particularly critical in dynamic scenarios. In such environments, the observation quality of a tracked feature can fluctuate wildly from one frame to the next. If the front end fails to filter out these unreliable observations, the errors inevitably bleed into the backend. Once they contaminate the state estimation and the marginalization process, the long-term accuracy of the system gradually degrades.

B. VIO/SLAM in Dynamic Environments

Navigating dynamic environments remains a central challenge for VIO and SLAM systems. Initial efforts leaned heavily on pure vision. Systems like DS-SLAM [23], DynaSLAM [4], DynaSLAM II [24], and SOF-SLAM [25] use semantic segmentation and motion consistency to detect and mask dynamic regions before they can corrupt pose estimation. Taking this a step further, object-level methods such as visual-inertial multi-instance SLAM [26] attempt to jointly model camera motion alongside moving objects and scene structure. Clearly, blending semantic and geometric cues helps stabilize localization when the environment is moving.

However, a key issue lies in how dynamic observations are handled. Most frameworks rely on frame-by-frame judgments. They classify a visual observation as either valid or invalid based on an instantaneous snapshot—such as a single

segmentation mask or immediate motion inconsistency. This works well when dynamic evidence is obvious. Yet, it quickly becomes brittle. Segmentation masks can be noisy. Motion is sometimes intermittent, and dynamic objects often stop moving temporarily. In aerial scenarios, these problems are amplified. Rapid viewpoint shifts and motion blur cause the quality of observations to swing unpredictably between consecutive frames.

To ground these fluctuating visual signals, researchers increasingly turn to inertial priors. By explicitly fusing IMU data with semantic and geometric cues, systems like DynaVINS [5] and D-VINS [6] build stronger defenses against dynamic interference. Extensions of this idea tackle specific edge cases: VINS-Dimc [27] applies multiple constraints, SRVIO [28] targets loop closures in moving scenes, and RD-VIO [29] uses IMU-informed matching to verify motion types. GMS-VINS [30] also shows the value of adapting semantic processing on the fly. Collectively, this body of work confirms that inertial data is essential for cross-checking visual dynamics.

But the front end is only half the battle. Long-term accuracy hinges heavily on back-end consistency. DM-VIO [7] illustrates this perfectly: how a system handles marginalization directly dictates its resilience over time. If the front end fails to adequately suppress a moving feature, that corrupted observation sneaks into the sliding-window optimization. Once it gets baked into the historical prior, the error becomes permanent, slowly degrading the entire trajectory.

Several limitations remain in current approaches. First, we still rely too much on short-sighted, frame-level suppression, leaving estimators vulnerable to transient errors like segmentation flickering. Furthermore, observation reliability is rarely tracked over time. This is especially true for heterogeneous data like point-line combinations. As noted earlier, partially corrupted lines are often handled clumsily—either kept entirely or thrown away, sacrificing useful geometric fragments. Finally, there is a systemic failure to protect the marginalized prior. While many algorithms down-weight bad features in the current optimization window, few provide a structural mechanism to stop these unreliable observations from polluting historical estimates.

These observations suggest that a more unified strategy is needed. To truly handle aggressive dynamic environments, the system needs a temporally consistent way to measure observation reliability. More importantly, this reliability metric must be woven seamlessly through both front-end feature association and back-end marginalization.

III. System Overview

The proposed framework focuses on improving the robustness of visual-inertial estimation in dynamic and low-texture environments, where unreliable visual observations may pass through the front end and affect back-end optimization. We break away from the standard practice of frame-wise binary feature rejection. Instead, we introduce a unified reliability model. This model evaluates the quality of each point and line over time, and the resulting reliability is used throughout the estimation pipeline. Fig. 1 outlines the overall architecture.

The pipeline ingests three primary streams of data, which are raw images, high-frequency IMU measurements, and semantic segmentation masks. Each plays a distinct role. Images yield the base point and line features. The IMU provides short-term kinematic predictions, serving as a physical baseline for geometric consistency checks. Meanwhile, the semantic masks supply category-level priors to flag potentially moving objects. Rather than processing these inputs in isolation, the system fuses them to dynamically estimate a continuous reliability score for each visual feature.

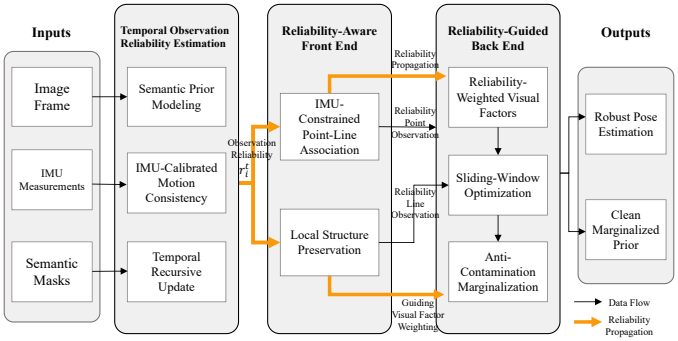


Fig. 1. Overall framework of the proposed method. Temporal observation reliability is estimated using semantic and inertial cues, and is further utilized in both front-end association and back-end marginalization.

We build upon a standard sliding-window visual-inertial formulation, tracking the vehicle's orientation, position, velocity, and sensor biases at each keyframe. Where our approach diverges is in the front-end processing. As points and lines are extracted, they are not just blindly tracked. They are assigned a temporally updated reliability score. We compute this by cross-referencing semantic priors with IMU-driven geometric verification, smoothing the results recursively over time.

This metric directly steers the front-end data association. Highly reliable features are actively preserved to anchor the local structure. Conversely, suspicious observations are suppressed before they can trigger faulty matches. Notably, this continuous scoring allows the system to salvage partially corrupted line features, preserving their valid segments rather than simply discarding the entire line.

The reliability-aware philosophy extends naturally into the back end. During optimization, the same confidence scores modulate the visual factors. Trustworthy features dominate the objective function. Uncertain ones are heavily down-weighted, or outright pruned, long before they can be absorbed into the marginalized prior. By stopping transient errors from accumulating within the sliding window, the system maintains stable, long-term consistency, even during aggressive maneuvers typical in UAV flight.

IV. Proposed Method

A. Temporal Observation Reliability Estimation for Aggressive Dynamic Scenarios

In aggressive UAV scenarios, both semantic predictions and geometric observations tend to become unstable. Semantic segmentation may produce incorrect labels or low-confidence outputs. At the same time, geometric measurements are easily affected by motion blur, rapid rotations, and depth ambiguity. These factors often appear together. As a result, decisions

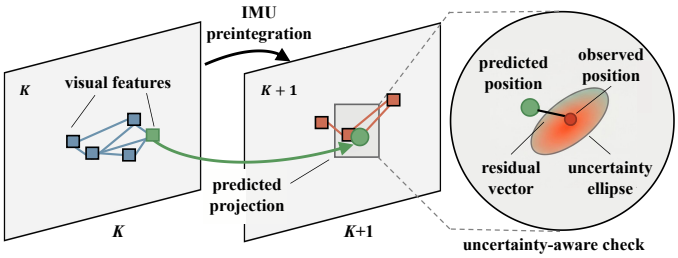


Fig. 2. IMU-calibrated kinematic consistency. A feature from the previous frame is projected to the current frame using IMU-preintegrated motion, and the difference between predicted and observed positions is evaluated with uncertainty.

made independently at each frame can be unreliable. Rather than forcing observations into binary categories, we treat the problem from a temporal perspective. Each observation is assigned with a continuous reliability value, which evolves over time. This value is not determined by a single cue. Instead, it is gradually refined by combining semantic priors, inertial consistency, and temporal recursion. This formulation can make the estimation less sensitive to short-term disturbances.

1) Semantic prior modeling: To introduce semantic cues without adding significant overhead, a lightweight YOLOv8n-seg model [31], pretrained on the MS COCO dataset [32], is employed. The semantic output is not used for direct filtering. Instead, it provides an initial estimate of observation reliability.

For each observation $\mathbf{u}_i^t$, the network outputs a category label c_i^t and a confidence score q_i^t. Each category is associated with a predefined coefficient $\rho_c \in [0,1]$, reflecting how likely it is to remain static. Static structures such as buildings are assigned higher values, while potentially dynamic objects are assigned lower ones. The semantic prior is then computed as Eq. (1):

$$s_i^t = \rho_{c_i^t} q_i^t + (1 - q_i^t) s_0 \tag{1}$$

where, s_0 is a neutral prior, typically set to 0.5. When the confidence is high, the model relies more on the semantic prediction. When it is low, the estimate falls back toward a neutral value. Compared with binary masking, this formulation keeps uncertainty in the representation. It also provides a smoother starting point for later updates.

2) IMU-calibrated kinematic consistency: Semantic information alone is often insufficient to determine whether an observation is truly static. To complement it, an IMU-based consistency check is introduced. The core idea is to evaluate whether a visual observation follows the rigid-body motion predicted by inertial measurements.

As illustrated in Fig. 2, IMU measurements between two consecutive frames are first preintegrated to estimate relative motion. A feature observed in the previous frame is then projected into the current frame [see Eq. (2)]:

$$\hat{\mathbf{u}}_i^t = \pi\left(\mathbf{T}_{cb}\hat{\mathbf{T}}_{t\leftarrow t-1}^{\text{imu}}\mathbf{T}_{bc}\mathbf{P}_i^{t-1}\right) \tag{2}$$

The difference between the predicted and observed positions forms the reprojection residual. Instead of applying a fixed threshold, this residual is evaluated with respect to its uncertainty. IMU noise, depth uncertainty, and image noise are all taken into account. The consistency is measured using the Mahalanobis distance [see Eq. (3)]:

$$m_i^t = (\mathbf{e}_i^t)^\top (\mathbf{\Sigma}_i^t)^{-1} \mathbf{e}_i^t \tag{3}$$

This allows the evaluation to adapt to different motion conditions. When motion is aggressive, uncertainty increases and the constraint becomes more tolerant. Under stable conditions, the check becomes stricter. In this way, the measure remains consistent with the underlying physical model.

3) Temporal recursion and reliability update: To reduce the impact of transient errors and occasional inconsistencies, reliability is updated over time rather than determined from a single frame. Each observation is treated as a temporal state. Three conditions are considered static, quasi-static, and dynamic. At each time step, the predicted state from the previous frame is combined with current evidence. This is done through a Bayesian-style update [see Eq. (4)]:

$$\mathbf{p}_i^t = \frac{\mathbf{l}_i^t \odot \bar{\mathbf{p}}_i^t}{\mathbf{1}^\top (\mathbf{l}_i^t \odot \bar{\mathbf{p}}_i^t)} \tag{4}$$

The update blends historical information with new observations. Short-term noise is less likely to dominate the result. The estimation therefore becomes more stable over time. A scalar reliability value is derived as Eq. (5):

$$r_i^t = P(z_i^t = S) + \eta P(z_i^t = Q) \tag{5}$$

where, $\eta \in [0,1]$ controls how much partially reliable observations contribute. The estimated reliability is used throughout the system. In the front end, it influences feature selection and association. In the back end, it is incorporated into visual factors and marginalization.

B. Reliability-Aware Point-Line Association and Local Structure Preservation

This module transforms the estimated observation reliability into robust front-end constraints. In aggressive UAV scenarios, rapid motion, viewpoint changes, and weak texture often reduce the stability of point features. Line features can provide complementary structural information, but they are sensitive to partial occlusion and local dynamic interference. To address these challenges, the proposed method integrates IMU-guided search, reliability-aware association, and local structure preservation to retain valid observations while suppressing unreliable ones.

1) IMU-constrained adaptive matching for point and line features: The front-end operates feature extraction based on ORB points and LSD line. For point features, the expected position in the current frame is predicted using IMU-preintegrated motion. Matching is then performed within a local search region centered on this prediction, rather than globally. The size of this region is adjusted based on motion uncertainty. Under stable motion conditions, the region is kept small to maintain efficiency and matching precision.

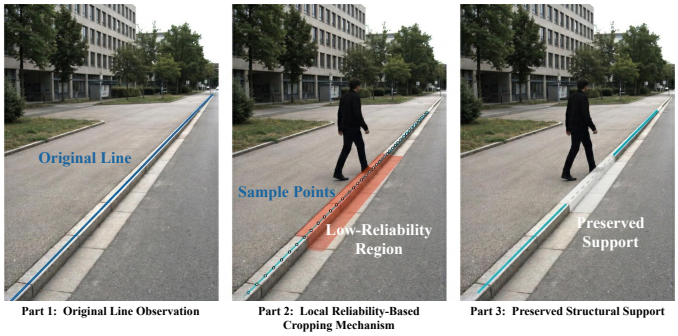


Fig. 3. Local line cropping and structure preservation mechanism. The reliability map is sampled along each matched line segment, and only the connected subregion with sufficiently high reliability is retained as the valid structural support.

During aggressive motion, the region is expanded to prevent the loss of valid correspondences. Line features are handled with a similar predictive approach. Each line segment from the previous frame is projected into the current frame, and candidate matches are searched within an uncertainty-aware band around the predicted location. This limits mismatches caused by viewpoint changes and assists tracking in low-texture scenes.

Observation reliability is also incorporated into the matching cost. For line features, the cost function is defined as Eq. (6):

$$C_k = \lambda_d d_{\mathrm{LBD}} + \lambda_\theta d_\theta + \lambda_\perp d_\perp + \lambda_r (1 - r^t_{\ell_c}) \tag{6}$$

The first three terms represent descriptor similarity and geometric consistency, while the final term applies a penalty based on reliability. As a result, matches exhibiting both geometric consistency and temporal reliability are prioritized. A similar reliability-based filtering mechanism is applied to point features to select stable candidates.

2) Local line cropping and structure preservation: Most existing methods treat line segments as complete entities, discarding the entire line if a portion overlaps with a dynamic region. This strategy can be overly restrictive in dynamic environments where lines may only be partially affected. The proposed local line cropping mechanism is illustrated in Fig. 3.

Instead of binary rejection, the proposed method evaluates lines locally. A reliability map is constructed on the image plane, and reliability values are sampled along the line segment. The average reliability is computed as Eq. (7):

$$r^t_\ell = \frac{1}{N} \sum_{s=1}^{N} r(\mathbf{x}_s) \tag{7}$$

While this average provides an indication of overall quality, it is not used directly for filtering. The line is parameterized, and sub-regions are evaluated individually. Only the segments with reliability exceeding a defined threshold are retained [see Eq. (8)]:

$$\Omega_{\mathrm{stat}} = \{ s \in [0,1] \mid r(\ell(s)) > \tau_{\mathrm{keep}} \} \tag{8}$$

Among the retained sections, the longest continuous segment is selected as the valid observation, and the remainder is removed. This makes it possible to preserve structural information even if the original line is partially degraded.

3) Joint reliability-aware optimization: After data association, observation reliability is further integrated into the optimization stage. The contribution of each observation is weighted by both its geometric residual and its estimated reliability. For point features, the reprojection residual is defined as Eq. (9):

$$\mathbf{e}_{p,i} = \mathbf{u}_i - \hat{\mathbf{u}}_i \tag{9}$$

For line features, the residual is computed using the preserved cropped segments, limiting the influence of unreliable portions. The joint optimization problem is formulated as Eq. (10):

$$\min_{\mathcal{X}} \sum_{i \in \mathcal{P}} w_{p,i} \rho \left(\|\mathbf{e}_{p,i}\|^2_{\mathbf{\Sigma}^{-1}_{p,i}} \right) + \sum_{k \in \mathcal{L}} w_{l,k} \rho \left(\|\mathbf{e}_{l,k}\|^2_{\mathbf{\Sigma}^{-1}_{l,k}} \right) \tag{10}$$

The $w_{p,i}$ and $w_{l,k}$ weights are determined by the temporal reliability. Consistent observations receive higher weights, while less reliable measurements are down-weighted.Robust kernels ρ are maintained to filter instantaneous outliers. The reliability weighting complements the robust kernel by reflecting the historical stability of the observation, which adds another way to handle dynamic interference during sliding-window estimation.

C. Reliability-Guided Anti-Contamination Marginalization

In sliding-window visual-inertial odometry, marginalization is used to keep the state size bounded and maintain computational efficiency. In dynamic or aggressive scenarios, however, some unreliable visual observations may pass front-end checks and enter the marginalized prior. Once included, these corrupted constraints can accumulate and lead to estimation drift over time. A reliability-guided marginalization strategy is introduced to reduce this effect. Observation reliability estimated in the front end is propagated to the prior construction stage, where visual factors with low reliability are down-weighted before marginalization. The overall pipeline is shown in Fig. 4.

Let $r_j \in [0,1]$ denote the reliability of the j-th visual factor. For point features, this value comes from temporal reliability estimation. For line features, it is derived from the retained line segment. Rather than using a hard threshold, a continuous gating function is applied to control the contribution of each factor [see Eq. (11)]:

$$\Gamma(r_j) = \begin{cases} 0, & r_j \le \tau_{\mathrm{drop}}, \\ \dfrac{1}{1 + \exp\left[-k\left(r_j - \tau_{\mathrm{marg}}\right)\right]}, & \tau_{\mathrm{drop}} < r_j < \tau_{\mathrm{safe}}, \\ 1, & r_j \ge \tau_{\mathrm{safe}}. \end{cases} \tag{11}$$

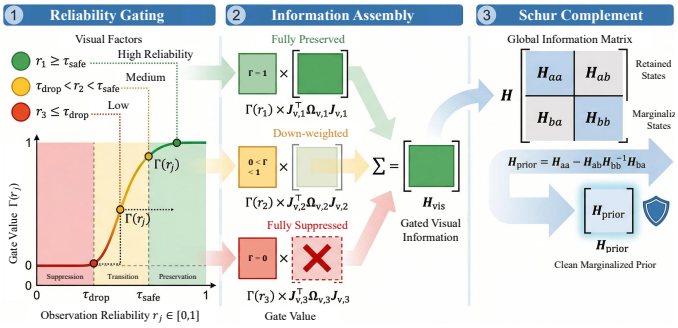


Fig. 4. Reliability-guided anti-contamination marginalization strategy. Observation reliability estimated in the front end is propagated to prior construction, so that suspicious visual constraints are weakened or excluded before being absorbed into the marginalized prior.

This function separates factors into three cases, low-reliability factors are removed, high-reliability ones are kept, and intermediate cases are smoothly weighted. Compared with binary selection, this formulation reduces sensitivity near the threshold. The gating is applied during linearization by scaling each visual factor. The visual information matrix becomes Eq. (12):

$$\mathbf{H}_{\text{vis}} = \sum_{j} \Gamma(r_j)\, \mathbf{J}_{v,j}^{\top} \mathbf{\Omega}_{v,j} \mathbf{J}_{v,j} \tag{12}$$

In this way, reliability is incorporated directly into the information matrix. Visual factors with low reliability contribute less to the optimization. After assembling the gated system, marginalization is carried out using the standard Schur complement. The state increment is partitioned into marginalized variables $\delta \mathbf{X}b$ and retained variables $\delta \mathbf{X}a$, yielding the prior [see Eq. (13)]:

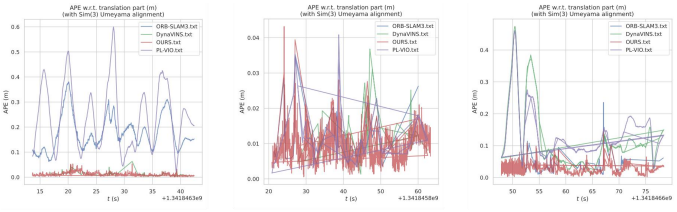
$$\mathbf{H}_{\text{prior}} = \mathbf{H}_{aa} - \mathbf{H}_{ab}\mathbf{H}_{bb}^{-1}\mathbf{H}_{ba} \tag{13}$$

Since unreliable factors are suppressed beforehand, the resulting prior mainly reflects reliable visual constraints and inertial information. A small damping term can be added if needed to improve numerical stability. Compared with conventional approaches, this method introduces a reliability-based filtering step before marginalization. This helps limit the influence of unreliable observations on the prior and improves consistency in dynamic environments.

V. Experimental Results

To evaluate the proposed method under different motion conditions and scene complexities, experiments are conducted on the EuRoC MAV benchmark and a self-collected UAV dataset. Both datasets provide synchronized image and IMU measurements, allowing the proposed framework to be evaluated as a complete visual-inertial odometry system. The evaluation focuses on scenarios involving rapid motion, weak texture, and dynamic interference.

The method is compared with several representative approaches, including ORB-SLAM3, DynaVINS, and PL-VIO.



(a) MH03 (b) V102 (c) V203

Fig. 5. APE of ORB-SLAM3, DynaVINS, PL-VIO, and the proposed method on the EuRoC dataset.

TABLE I. Comparison of the Average RMSE [m] of the Proposed Method with ORB-SLAM3, DynaVINS, and PL-VIO on the EuRoC Dataset.

Sequence	ORB-SLAM3	DynaVINS	PL-VIO	Ours
MH01	0.075	0.084	0.210	0.062
MH02	0.044	0.105	0.232	0.037
MH03	0.217	0.054	0.312	0.046
MH04	0.132	0.122	0.139	0.075
MH05	0.121	0.147	0.257	0.057
V101	0.059	0.047	0.059	0.049
V102	0.023	0.018	0.016	0.011
V103	0.096	0.180	0.130	0.037
V201	0.040	0.056	0.032	0.042
V202	0.062	0.090	0.095	0.021
V203	0.104	0.154	0.201	0.027
Average	0.088	0.096	0.153	0.042

The evaluation focuses on scenarios involving dynamic interference, weak texture, and large motion. For quantitative analysis, Absolute Pose Error (APE) and Absolute Trajectory Error (ATE) are used as the primary metrics. APE reflects the overall trajectory deviation, while ATE focuses on translational accuracy. RMSE is also reported to give a more compact summary and to capture sensitivity to larger errors.

A. Evaluation on the EuRoC Dataset

The EuRoC MAV dataset [33] is widely used in visual-inertial odometry research. It provides synchronized stereo images and IMU measurements, together with ground-truth trajectories. The dataset includes two indoor environments, Machine Hall (MH) and Vicon Room (V), covering motion patterns from relatively stable to aggressive.

All 11 sequences are included in the evaluation, namely MH01-MH05, V101-V103, and V201-V203. These sequences involve various challenges, such as rapid motion, motion blur, repeated structures, and weak-texture regions. Some sequences are particularly demanding, which makes them suitable for testing robustness.

As can be seen in Fig. 5, the proposed method generally produces lower errors with reduced fluctuations compared to the baseline methods. On relatively stable sequences such as V102, the performance of all methods is quite close. On more challenging sequences like MH03 and V203, fewer error spikes can be observed, and the estimates remain more stable overall. This behavior is likely related to the temporal reliability mechanism, which helps mitigate the impact of degraded observations.

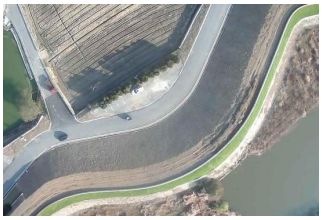
From Table I, the proposed method achieves an average RMSE of 0.042 m, which is lower than ORB-SLAM3, Dy-



(a) TR9S Hexacopter UAV



(b) TR9S lands around the lake



(c) Urban road data scene



(d) Forest and field data scene

Fig. 6. The data collection system and typical test scenes.

naVINS, and PL-VIO. The improvement becomes more noticeable on challenging sequences such as MH03, MH05, V103, V202, and V203, where motion complexity and dynamic interference are more pronounced. On sequences such as V101 and V201, the proposed method does not obtain the lowest error, although the differences remain relatively small. Taken together, the results suggest that the method maintains stable performance across different sequences and shows improved robustness under varying visual-inertial conditions.

B. Evaluation on Self-Collected UAV Flight Dataset

To further examine the method under real flight conditions, we collected a UAV dataset using a hexarotor platform equipped with a monocular camera, an IMU, and a differential GPS/RTK system. The camera and IMU data are used for odometry estimation, while the RTK measurements provide ground truth. The camera runs at 30Hz and the IMU at 100Hz. All sensors are time-synchronized, and the camera–IMU extrinsics are calibrated offline. For consistency, all baseline methods use the same input data, and trajectory errors are computed against the RTK reference.

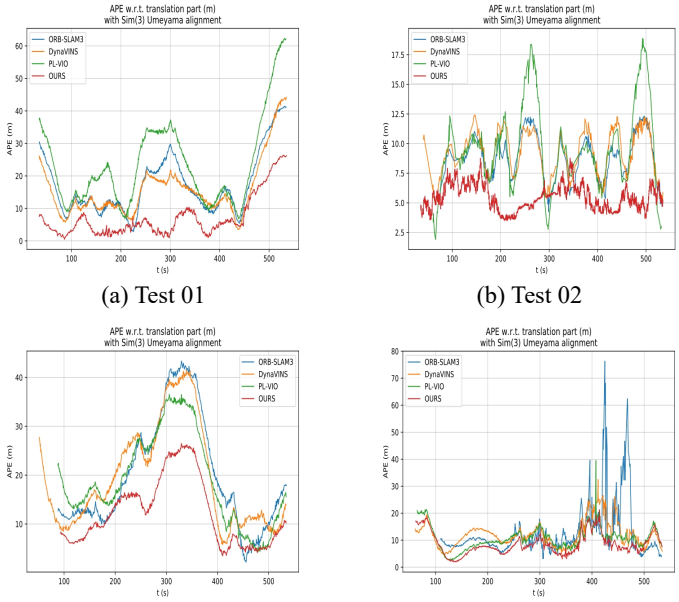
Five sequences, denoted as Test01-Test05, are recorded in different environments, including urban areas, lakes, forests, and suburban roads. The total flight distance is about 20km. These sequences include a range of conditions such as aggressive motion, weak-texture regions, and dynamic objects like vehicles and pedestrians. The data collection platform and representative test scenes are shown in Fig. 6.

Trajectory evaluation is carried out using the EVO toolkit after aligning the estimated trajectories with the RTK ground truth.

From Table II, the proposed method achieves an average RMSE of 9.51 m, which is lower than ORB-SLAM3, DynaVINS, and PL-VIO. The difference is more visible on sequences such as Test01, Test04, and Test05, where longer trajectories, stronger motion, and dynamic interference make estimation more difficult.

TABLE II. COMPARISON OF RMSE [m] OF THE PROPOSED METHOD WITH ORB-SLAM3, DYNAVINS, AND PL-VIO ON THE UAV DATASET.

Sequence	ORB-SLAM3	DynaVINS	PL-VIO	Ours
Test01	30.52	17.79	27.41	8.45
Test02	8.75	9.65	10.80	5.42
Test03	10.56	10.95	42.68	9.87
Test04	36.09	23.25	22.29	14.55
Test05	46.14	12.79	11.46	9.26
Average	26.41	14.89	22.93	9.51



(a) Test 01 (b) Test 02

(c) Test 04 (d) Test 05

Fig. 7. APE curves of ORB-SLAM3, DynaVINS, PL-VIO, and the proposed method on the self-collected UAV dataset.

As seen in Fig. 7, the proposed method generally maintains lower errors and smoother trends along the trajectories. The baseline methods show larger fluctuations in several segments, together with occasional spikes, which suggests local tracking instability.

This difference becomes clearer in segments with rapid motion or degraded visual input. In such cases, ORB-SLAM3 and DynaVINS tend to produce noticeable error increases and drift over time, while the proposed method remains more stable and avoids abrupt degradation. This behavior indicates that the reliability-aware design can help improve consistency over longer trajectories in dynamic outdoor scenes.

C. Ablation Study

To further analyze the contribution of the proposed components, two ablation studies are conducted on the self-collected UAV dataset. The reported RMSE values are averaged over the five UAV sequences. These sequences include outdoor scenes with weak textures, camera motion, and dynamic interference, and therefore provide a practical basis for evaluating the stability of each module.

Table III presents the ablation results of the reliability-aware VIO framework on the challenging UAV sequences.

TABLE III. ABLATION STUDY OF THE PROPOSED RELIABILITY-AWARE VIO FRAMEWORK.

Temporal Observation Reliability Estimation	Point-Line Association and Local Structure Preservation	Reliability-Guided Marginalization	RMSE (m)	Relative percentage Reduction in RMSE
-	-	-	22.93	-
✓	-	-	15.09	34.19%
-	✓	-	17.64	23.07%
-	-	✓	13.33	41.87%
✓	✓	✓	9.51	58.53%

TABLE IV. ABLATION STUDY OF THE TEMPORAL RELIABILITY ESTIMATION MODULE.

Semantic Prior	IMU-Based Motion Consistency	RMSE (m)	Relative percentage reduction in RMSE
-	-	20.39	-
✓	-	16.58	18.69%
-	✓	15.24	25.26%
✓	✓	10.75	47.28%

Each module brings a certain improvement over the base point-line VIO system. Temporal observation reliability estimation reduces the average RMSE from 22.93 m to 15.09 m, which suggests that modeling the credibility of visual observations over time helps suppress unstable measurements. The point-line association and local structure preservation strategy also improves the result by keeping useful geometric information from partially affected line features. Reliability-guided marginalization gives the largest individual reduction among the three components. Preventing unreliable visual constraints from being absorbed into the marginalized prior is helpful for limiting accumulated drift. With all three components enabled, TRAIL-VIO achieves the lowest average RMSE of 9.51 m.

Table IV reports the average performance of the temporal reliability estimation module across the five UAV sequences. Using semantic priors alone improves the RMSE from 20.39 m to 16.58 m, mainly because potentially dynamic regions can be identified before they strongly affect pose estimation. IMU-based motion consistency alone leads to a slightly lower RMSE of 15.24 m, showing that inertial prediction provides a useful motion-level check when visual observations become uncertain. The combination of both cues achieves the best average result. Semantic information and IMU-based consistency are not simply redundant, and they tend to compensate for each other in temporal reliability estimation.

D. Parameter Sensitivity Analysis

To address potential concerns regarding the generalizability of the manually configured parameters and to evaluate the robustness of the proposed framework, a comprehensive sensitivity analysis is conducted. We select two highly representative and challenging sequences for this evaluation: MH05 from the EuRoC dataset, which features indoor aggressive motion and low-illumination, and Test04 from the self-collected UAV dataset, which encompasses long-distance outdoor flight with strong dynamic interference. The APE RMSE is employed as the primary evaluation metric.

1) Sensitivity to Reliability Gating Thresholds: The gating parameters τ_{drop} and τ_{safe} control the reliability-guided marginalization strategy in Eq. (11), where visual factors are divided into suppression, transition, and preservation regions according to their estimated reliability. To examine their individual effects, a one-factor-at-a-time control-variable strategy is adopted. In each test, only one threshold is varied, while the remaining parameters are kept fixed at their default values.

TABLE V. SENSITIVITY OF APE RMSE [m] TO THE DROP THRESHOLD τ_{drop}.

τ_{drop}	MH05	Test04
0.00	0.082	18.20
0.05	0.071	16.45
0.10	0.063	15.10
0.15	0.058	14.62
0.20	0.057	14.55
0.25	0.057	14.56
0.30	0.058	14.58
0.35	0.059	14.65
0.40	0.064	15.30
0.45	0.076	17.85
0.50	0.095	21.40

TABLE VI. SENSITIVITY OF APE RMSE [m] TO THE SAFE THRESHOLD τ_{safe}.

τ_{safe}	MH05	Test04
0.50	0.065	15.80
0.55	0.062	15.30
0.60	0.060	14.95
0.65	0.059	14.70
0.70	0.058	14.58
0.75	0.057	14.56
0.80	0.057	14.55
0.85	0.057	14.55
0.90	0.058	14.57
0.95	0.059	14.62
1.00	0.063	15.15

First, the influence of the lower suppression threshold τ_{drop} is evaluated. τ_{safe} and the steepness factor k are fixed, and τ_{drop} is varied from 0.0 to 0.5. The results are reported in Table V. When τ_{drop} is close to 0.0, the suppression effect becomes weak. As a result, some unreliable observations caused by motion blur, dynamic objects, or inconsistent feature associations may still contribute to the optimization and the marginalized prior, leading to an increase in APE RMSE. In contrast, when τ_{drop} is set too high, especially above 0.45, the system tends to suppress more visual factors than necessary. This may remove valid geometric constraints and weaken the visual support for pose estimation. A relatively stable region can be observed when τ_{drop} lies in the interval $[0.15, 0.35]$. Within this range, the RMSE values remain close to the baseline results, approximately $0.057\,\mathrm{m}$ on MH05 and $14.55\,\mathrm{m}$ on Test04.

A similar analysis is conducted for the upper preservation threshold τ_{safe}. In this experiment, τ_{drop} is fixed at its default value, and τ_{safe} is swept from 0.5 to 1.0. As shown in Table VI, stable performance can be observed within a relatively broad interval, and the performance does not collapse under moderate parameter perturbations. Therefore, TRAIL-VIO does not depend on a narrowly selected pair of gating parameters. The default values of τ_{drop} and τ_{safe} are located within stable performance regions rather than at isolated optimal points.

2) Sensitivity to the Transition Steepness Factor: After analyzing the two reliability gating thresholds, we further evaluate the effect of the transition steepness factor k. The steepness factor k controls the shape of the sigmoid weighting function in Eq. (11). It determines how rapidly the reliability

TABLE VII. SENSITIVITY OF APE RMSE [m] TO THE TRANSITION STEEPNESS FACTOR k.

k	MH05	Test04
1	0.063	15.12
5	0.058	14.65
10	0.057	14.55
15	0.057	14.59
20	0.059	14.72
50	0.065	15.38

weight changes within the transition region between τ_{drop} and τ_{safe}. A small k produces a smoother transition, while a large k makes the weighting function closer to a hard switch. Therefore, this parameter affects how gently or aggressively visual factors are down-weighted when their reliability values fall between the suppression and preservation thresholds. In this experiment, τ_{drop} and τ_{safe} are fixed at their default values, and only k is varied. The tested values are selected as $k = \{1, 5, 10, 15, 20, 50\}$, covering a wide range from nearly linear transition to a steep, almost binary gating behavior.

The results in Table VII show that a moderate steepness factor generally leads to more stable performance. Setting k to an extremely low value like 1 tends to under-penalize questionable observations. The transition becomes too flat. Consequently, some degraded features retain a disproportionate amount of influence during the information matrix assembly, which causes a slight increase in the tracking error. And pushing k to 50 mimics a rigid binary threshold. This sharp cutoff can abruptly discard features that are only experiencing transient ambiguity. Such sudden losses of structural constraints occasionally introduce minor instabilities into the sliding-window optimization.

The proposed framework demonstrates considerable tolerance between these two extremes. For values of k roughly between 5 and 15, the APE RMSE remains largely consistent. On the MH05 sequence, the error fluctuates by merely a few millimeters within this moderate range. The Test04 sequence shows a similarly stable plateau. This behavior implies that the soft-gating mechanism successfully bridges the gap between total suppression and full preservation. By smoothing the factor weighting, the system generally avoids overreacting to short-term observation noise.

3) Tolerance to Semantic Prior Perturbation: In addition to the reliability gating parameters, we further evaluate the influence of the semantic category coefficient ρ_c. In the proposed framework, ρ_c represents the prior static reliability of a semantic category in the semantic reliability model. Static structures, such as buildings, roads, and background regions, are assigned relatively high prior values, while potentially dynamic objects, such as pedestrians and vehicles, are assigned lower prior values. Since these coefficients are manually predefined, it is necessary to examine whether inaccurate semantic priors would cause a noticeable degradation in trajectory estimation.

To simulate possible miscalibration of semantic priors in unseen environments, a perturbation factor Δ_ρ is introduced to the default semantic coefficient. The perturbed coefficient is defined as Eq. (14):

$$\tilde{\rho}_c = \text{clip}(\rho_c + \Delta_\rho, 0, 1), \tag{14}$$

TABLE VIII. SENSITIVITY OF APE RMSE [m] TO SEMANTIC PRIOR PERTURBATION.

Perturbation Δ_ρ	MH05	Test04
-30%	0.066	15.32
-20%	0.062	14.95
-10%	0.059	14.68
0	0.057	14.55
$+10\%$	0.058	14.65
$+20\%$	0.061	14.88
$+30\%$	0.065	15.25

where, $\tilde{\rho}_c$ denotes the perturbed semantic coefficient, and $\text{clip}(\cdot)$ limits the perturbed value to the valid interval $[0, 1]$. In this experiment, Δ_ρ is swept from -0.30 to $+0.30$ with a step size of 0.10. The case of $\Delta_\rho = 0$ represents the default semantic prior setting used in the main experiments.

During this analysis, all semantic categories are divided into two broad groups, static categories and potentially dynamic categories. The same perturbation strategy is applied to the semantic coefficients of these categories to evaluate the tolerance of the system to prior uncertainty. This setting is intentionally simple. It does not aim to find category-specific optimal weights. Instead, it tests whether the proposed reliability estimation framework remains stable when the manually assigned semantic priors are not fully accurate. The other parameters, including τ_{drop}, τ_{safe}, and k, are fixed at their default values.

The results are reported in Table VIII. When negative perturbations are introduced, the semantic reliability assigned to static categories decreases, and some stable background features may contribute less to the visual constraints. Positive perturbations have the opposite effect. They may increase the reliability of potentially dynamic or ambiguous regions, so these observations are less strongly penalized by the semantic prior. Both cases can affect the APE RMSE to some extent. Within the tested perturbation range, the variation remains limited, as the final observation reliability is still refined by IMU-based motion consistency and temporal reliability updating.

TRAIL-VIO is not overly sensitive to precisely tuned semantic category coefficients. Semantic information provides an initial prior, rather than a final reliability decision. The subsequent IMU-calibrated kinematic consistency and temporal update help reduce the influence of imperfect semantic priors. Therefore, the proposed framework shows a certain tolerance to manually assigned category weights and does not require highly precise semantic coefficient tuning for every new environment.

E. Runtime Analysis

To evaluate the computational overhead of TRAIL-VIO, the average runtime of the main modules is measured on a laptop equipped with an Intel Core i9-14900HX CPU, 31.1 GB RAM, and an NVIDIA GeForce RTX 4060 GPU under Ubuntu 18.04.6 LTS. The runtime is reported in milliseconds per frame and averaged on the self-collected UAV sequences.

We separate YOLOv8n-seg inference from temporal observation reliability estimation. It only includes the reliability assignment based on semantic outputs, IMU-based motion consistency evaluation, and temporal reliability updating.

TABLE IX. AVERAGE COMPUTATION TIME [ms] OF THE TRAIL-VIO SYSTEM ON SELF-COLLECTED UAV DATASET.

SEQUENCE	POINT AND LINE FEATURE EXTRACTION / TRACKING	YOLOV8N-SEG INFERENCE	TEMPORAL OBSERVATION RELIABILITY ESTIMATION	RELIABILITY-AWARE POINT-LINE ASSOCIATION AND LOCAL STRUCTURE PRESERVATION	BACK-END OPTIMIZATION AND RELIABILITY-GUIDED MARGINALIZATION	TOTAL RUNTIME
Test01	38.76	30.18	3.42	8.71	35.66	116.73
Test02	31.84	27.36	2.95	6.83	27.42	96.40
Test03	35.29	28.11	3.18	7.52	31.36	105.46
Test04	43.67	31.74	3.89	9.34	42.15	130.79
Test05	40.52	29.63	3.61	8.97	38.64	121.37
Average	38.02	29.40	3.41	8.27	35.05	114.15

As shown in Table IX, the average total runtime of TRAIL-VIO is 114.15 ms per frame, corresponding to approximately 8.76 FPS. The runtime varies across different UAV sequences. This variation is mainly related to the number of extracted and tracked point-line features, the density of valid residuals, keyframe insertion, and the scale of sliding-window optimization.

The current implementation has not yet been fully optimized for strict onboard real-time deployment. Further acceleration may be achieved by TensorRT-based semantic inference, reducing the frequency of semantic updates, optimizing line feature extraction, and adopting an asynchronous front-end pipeline.

VI. CONCLUSION

This study presents TRAIL-VIO, a reliability-aware point-line visual-inertial odometry framework for dynamic and low-texture environments. The proposed method models visual observation quality as a temporally updated reliability state rather than making frame-wise binary decisions. By combining semantic priors with IMU-based motion consistency, TRAIL-VIO suppresses transiently degraded observations caused by motion blur, occlusion, weak texture, and moving objects. The estimated reliability is further used in point-line association, local line-structure preservation, optimization, and marginalization, allowing the system to retain useful structural information while reducing the influence of unreliable visual constraints. Experiments on the EuRoC MAV dataset and the self-collected UAV dataset show that TRAIL-VIO generally achieves more accurate and stable trajectories than the compared methods, especially under rapid motion, dynamic interference, and weak visual texture. The additional analyses further support the role of the proposed modules and show that the framework remains reasonably stable under moderate parameter variations, with most of the extra runtime arising from semantic segmentation. Despite these improvements, several practical limitations remain. The method still depends on semantic segmentation quality. Although the sensitivity analysis shows that the framework can tolerate perturbations of up to $\pm 30\%$ in the initial semantic priors, persistent semantic errors may still degrade the estimation, especially when dynamic objects are repeatedly assigned high reliability. The reliability-related parameters are also manually configured. The system maintains relatively stable tracking when $\tau_{\mathrm{drop}} \in [0.15, 0.35]$ and $k \in [5, 15]$, but these ranges may shift for sensors with different noise characteristics, frame rates, or motion patterns. In addition, the point-line front end still requires sufficient valid geometric constraints. In structure-poor scenes, especially under sustained aggressive rotations, the loss of reliable visual updates may leave IMU preintegration insufficiently corrected and lead to rapid drift. Future work will focus on adaptive reliability estimation, efficient semantic inference, and tighter visual-inertial coupling for more robust onboard deployment.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) Joint Fund of Civil Aviation Research under grant U2333214, the Sichuan Science and Technology Program under grant 2026YFHZ0275, the Sichuan Flight Engineering Technology Research Center Project under grant GY2024-11C and the Civil Aviation Administration of China Safety Capacity Building Project under grant MHAQ2024033.

REFERENCES

[1] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multi-map SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[3] Y. He, J. Zhao, Y. Guo, W. He, and K. Yuan, "PL-VIO: Tightly-coupled monocular visual-inertial odometry using point and line features," *Sensors*, vol. 18, no. 4, p. 1159, 2018.

[4] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, 2018.

[5] S. Song, H. Lim, A. J. Lee, and H. Myung, "DynaVINS: A visual-inertial SLAM for dynamic environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11523–11530, 2022.

[6] Y. Sun, Q. Wang, C. Yan, Y. Feng, R. Tan, X. Shi, and X. Wang, "D-VINS: Dynamic adaptive visual-inertial SLAM with IMU prior and semantic constraints in dynamic scenes," *Remote Sens.*, vol. 15, no. 15, p. 3881, 2023.

[7] L. von Stumberg and D. Cremers, "DM-VIO: Delayed marginalization visual-inertial odometry," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1408–1415, 2022.

[8] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2007, pp. 3565–3572.

[9] K. Sun, B. Mohta, K. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 965–972, 2018.

[10] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 298–304.

[11] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial SLAM using nonlinear optimization," in *Proc. Robot. Sci. Syst. (RSS)*, 2013.

[12] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020, pp. 4666–4672.

[13] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," *arXiv preprint arXiv:1901.03638*, 2019.

[14] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "PL-SLAM: Real-time monocular visual SLAM with points and lines," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 4503–4508.

[15] Q. Fu, J. Wang, H. Yu, I. Ali, F. Guo, Y. He, and H. Zhang, "PL-VINS: Real-time monocular visual-inertial SLAM with point and line features," *arXiv preprint arXiv:2009.07462*, 2020.

[16] F. Zheng, G. Tsai, Z. Zhang, S. Liu, C.-C. Chu, and H. Hu, "Trifo-VIO: Robust and efficient stereo visual inertial odometry using points and lines," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 3686–3693.

[17] Z. Zhao, T. Song, B. Xing, Y. Lei, and Z. Wang, "PLI-VINS: Visual-inertial SLAM based on point-line feature fusion in indoor environment," *Sensors*, vol. 22, no. 14, p. 5457, 2022.

[18] F. Zheng, L. Zhou, W. Lin, J. Liu, and L. Sun, "LRPL-VIO: A lightweight and robust visual-inertial odometry with point and line features," *Sensors*, vol. 24, no. 4, p. 1322, 2024.

[19] J. He, M. Li, Y. Wang, and H. Wang, "PLE-SLAM: A visual-inertial SLAM based on point-line features and efficient IMU initialization," *IEEE Sensors J.*, vol. 25, no. 4, pp. 6801–6811, 2025.

[20] Y. Zhang, P. Zhu, and W. Ren, "PL-CVIO: Point-line cooperative visual-inertial odometry," in *Proc. IEEE Conf. Control Technol. Appl. (CCTA)*, 2023, pp. 859–865.

[21] D. Zou, Y. Wu, L. Pei, H. Ling, and W. Yu, "StructVIO: Visual-inertial odometry with structural regularity of man-made environments," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 999–1013, 2019.

[22] S. Yang, Y. Song, M. Kaess, and S. Scherer, "Pop-up SLAM: Semantic monocular plane SLAM for low-texture environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2016, pp. 1222–1229.

[23] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 1168–1174.

[24] B. Bescos, C. Campos, J. D. Tardós, and J. Neira, "DynaSLAM II: Tightly-coupled multi-object tracking and SLAM," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5191–5198, 2021.

[25] L. Cui and C. Ma, "SOF-SLAM: A semantic visual SLAM for dynamic environments," *IEEE Access*, vol. 7, pp. 166528–166539, 2019.

[26] Y. Ren, B. Xu, C. L. Choi, and S. Leutenegger, "Visual-inertial multi-instance dynamic SLAM with object-level relocalisation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2022, pp. 11055–11062.

[27] D. Fu, H. Xia, Y. Liu, and Y. Qiao, "VINS-DIMC: A visual-inertial navigation system for dynamic environment integrating multiple constraints," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 2, p. 95, 2022.

[28] A. Samadzadeh and A. Nickabadi, "SRVIO: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions," *IEEE Trans. Robot.*, vol. 39, no. 4, pp. 2878–2891, 2023.

[29] J. Li, X. Pan, G. Huang, Z. Zhang, N. Wang, H. Bao, and G. Zhang, "RD-VIO: Robust visual-inertial odometry for mobile augmented reality in dynamic environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 10, pp. 6941–6955, 2024.

[30] R. Zhou, J. Liu, J. Xie, J. Zhang, Y. Hu, and J. Zhao, "GMS-VINS: Multi-category dynamic objects semantic segmentation for enhanced visual-inertial odometry using a promptable foundation model," *arXiv preprint arXiv:2411.19289*, 2024.

[31] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub repository, version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[33] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

Trustworthy App Detection in Saudi Mobile Finance: Bridging Deep Learning and Interpretability

Raed Alharbi, Maryam Alghamdi
College of Computing and Informatics
Saudi Electronic University, Riyadh 11673, Saudi Arabia

***Abstract*—Understanding user trust in mobile financial applications is crucial as these platforms increasingly shape how users in Saudi Arabia manage finances and engage with digital banking. However, existing deep learning–based detection models often function as black boxes, offering limited interpretability, while traditional machine learning models, though more transparent, fail to capture complex interactions between permissions, reviews, and user behaviors. To address this gap, we propose Trustworthy App Detection for Saudi Arabia (TAD-Saudi) - a novel framework for interpretable and behavior-aware trust evaluation. The framework integrates the representational power of deep learning with the explainability of simpler models, enabling both global and local interpretation of trust-related features. Experimental results show that TAD-Saudi outperforms traditional baselines across multiple models. Moreover, the analysis reveals that users may continue to trust applications requesting sensitive permissions, particularly when these apps have high ratings or positive reviews.**

***Keywords*—XAI; trustworthy apps detection; encoding**

I. Introduction

The widespread adoption of financial applications in Saudi Arabia has transformed the way users manage their finances, perform transactions, and engage with digital banking services. With increasing reliance on mobile applications, the Saudi fin-tech sector has experienced significant growth, reflecting broader global trends in the adoption of mobile finance. A recent market analysis indicates a notable increase in financial app usage in the region, with a substantial increase in mobile banking transactions and fintech investments [1]. However, alongside this growth, concerns about app trustworthiness, data privacy, and security risks have emerged, particularly as users trust these applications with sensitive financial and personal data.

To develop an effective and trustworthy Saudi financial app detection model, it is crucial to: 1) understand the key factors that influence user trust in financial applications, such as app ratings, reviews, and security permissions; 2) analyze the role of such factors in the trustworthiness of an app, as excessive permission requests often raise security concerns among users; and 3) address the "black-box" nature of machine learning-based classification models [2], which often lack interpretability, making it difficult for users to rely on the classification outcomes with confidence. These factors emphasize the need for a detection framework that not only ensures high accuracy when designing machine learning models but also provides explainability and transparency to users.

Despite notable progress in app fraud detection, current detection methods lack explainability and user-centered transparency. Most existing approaches rely on sentiment analysis or metadata-driven classification, which focus on identifying fraudulent or malicious apps rather than explaining why an app is considered trustworthy. This absence of interpretability limits user confidence and delays the practical adoption of Artificial Intelligence-based (AI-based) trust evaluation systems. Furthermore, existing models often fail to capture the effect of user trust signals (such as ratings and reviews) and app behavioral features (such as permissions and update frequency). While deep learning models provide high accuracy, their decisions remain opaque, and traditional machine learning models, though interpretable, are too simplistic to capture these complex behavioral relationships.

To overcome these challenges, we propose TAD-Saudi, a framework that is designed to enhance the evaluation of mobile app trustworthiness. The core of TAD-Saudi consists of three main components. First, we introduce a novel benchmark that is designed to assess key cybersecurity features. This benchmark systematically collects and evaluates various factors that influence the acceptance or rejection of app trustworthiness, offering a robust foundation for understanding the critical elements that contribute to establishing trust in mobile applications. Second, TAD, our proof-of-concept framework, integrates the power of deep learning models with the explainability of simpler, interpretable models. Finally, we conduct an in-depth analysis to examine the underlying factors driving user trust, providing valuable insight into the behavioral aspects influencing app adoption.

The remainder of this study is organized as follows. Section II presents the related works. The proposed architecture is introduced in Section III, where Section IV provides the experiment setup, analysis, and discussion of our findings. Section VI concludes the study with a summary of our contributions.

II. Related Work

Several recent studies provide valuable insights into the evolving landscape of privacy concerns and trustworthiness assessment for mobile applications. Habib et al. [3] introduce Trust4App, a framework that automates mobile app trustworthiness assessment by combining public metrics (e.g., ratings, reviews) with security factors, and user-specific preferences. This model integrates existing approaches, such as static malware analysis, into a unified personalized trust score. Similarly, [4] propose a graph-based framework that assesses smartphone apps' trustworthiness by integrating ratings, reviews,

and permissions data into confidence metrics. While both frameworks contribute to multi-dimensional trust assessment, they rely primarily on static metadata and lack mechanisms for interpretability, limitations that motivate the explainable and behavior-aware design of TAD-Saudi.

Akgul et al. [5] conduct an analysis of privacy-related app reviews on the Google Play Store over ten years (2013–2023) using Natural Language Processing (NLP) techniques. Their findings reveal a growing shift in user attention from traditional permission-based privacy issues toward emerging risks such as data deletion, third-party sharing, and data theft, while also highlighting regional variations in user behavior and privacy expectations. Nema et al. [6] propose an NLP-based framework for extracting and categorizing privacy concerns from Android app reviews, identifying key issues such as unnecessary permissions, data collection practices, user tracking, and insufficient privacy controls. Although these studies provide valuable insights into user perceptions of privacy, they remain primarily descriptive and lack quantitative or interpretable detection mechanisms.

Recent efforts have highlighted the growing impact of explainable AI (XAI) in enhancing transparency and reliability within financial and cybersecurity frameworks. For example, Turgut et al. [7] employ XAI techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Explain Like I'm 5 (ELI5) to improve the interpretability of Machine Learning models used in financial fraud detection, emphasizing the importance of explainability for fostering user trust in automated decision systems. Similarly, Xu et al. [8] propose an interpretable credit-default assessment model that integrates multiple Machine Learning (ML) architectures with SHAP-based explanations, demonstrating how interpretability can enhance both trust and model comprehension in financial risk analysis.

Harkous et al. [9] propose Hark, a deep learning–based system that leverages NLP to analyze privacy-related app reviews, providing developers with actionable insights into user feedback and privacy perceptions. Similarly, [10] introduced MARS (Mobile App Reviews Summarization), an ML–driven framework that combines NLP and sentiment analysis techniques to automatically summarize user reviews from the Google Play Store, distilling key privacy concerns and recurring user issues. While both approaches enhance developers' understanding of user privacy feedback, they remain developer-centric and lack mechanisms for transparent or explainable trust assessment.

In the context of fraudulent mobile apps, Rani et al. [11] develop a sentiment analysis framework that detects fraudulent apps by analyzing user reviews and ratings to identify suspicious applications. Similarly, Zaki et al. [12] propose a robust machine learning architecture for detecting fraudulent mobile apps. Their findings demonstrate the higher accuracy of deep models, particularly Convolutional Neural Networks (CNNs), in detecting deceptive behavior and enhancing app security. However, these detection systems operate largely as black boxes, focusing on performance metrics without offering interpretable reasoning behind classification outcomes. TAD-Saudi diverges from these approaches by embedding explainability within its detection framework.

While there has been considerable progress in developing frameworks for assessing mobile app trustworthiness and detecting fraudulent applications, these approaches remain limited in several key aspects. Most existing frameworks focus primarily on classification accuracy without examining the underlying behaviors that drive trust predictions. They often fail to explore the factors influencing machine learning models' decisions, which is crucial for understanding why users trust certain apps and based on what criteria. Without this interpretability, users and developers lack meaningful insights into how security, privacy, and app behavior shape trustworthiness assessments.

III. The Proposed Architecture

This study introduces TAD-Saudi, as illustrated in Fig. 1, where the main components are 1) a novel benchmark that is designed to assess cybersecurity features, systematically collecting and evaluating various factors that influence the acceptance or rejection of app trustworthiness. This benchmark provides a comprehensive framework for understanding the underlying elements contributing to the establishment of trust in mobile applications; 2) TAD, a proof-of-concept framework that balances the richness of deep learning with the explainability of simpler models to predict and classify trustworthy apps, thus enabling the identification of secure and reliable mobile applications.

This section starts with an overview of the dataset, followed by detailed explanations of the proposed framework.

A. Data Preparation

Given the nature of this study, which investigates financial applications in the Saudi region, our focus is specifically on mobile apps that support the Arabic language. This emphasis is crucial, as the majority of users in the region rely on Arabic-language interfaces for their digital interactions.

We construct our dataset through a combination of web scraping techniques and APIs to gather comprehensive data on finance-related mobile applications from the Google Play Store [13]. Specifically, we utilize Python libraries such as $requests$, $BeautifulSoup$, and $google_play_scraper$ to extract detailed information on applications, including their descriptions, user reviews, ratings, and security permissions. To ensure a broad and representative coverage of financial applications, we collect data from 23 finance-related categories, such as "Banking," "Investment," "Loans," "Insurance," "Expense Tracking", and "Financial Planning" [13]. By focusing on these categories, we ensure a diverse dataset that specifically serves the needs of Arabic-speaking users in Saudi Arabia, reflecting the region's unique financial landscape and user preferences. To address the possibility of the Python libraries extracting irrelevant results, each extracted application was manually reviewed. During this process, all duplicate applications and any applications not directly related to financial services were removed. This ensured that the final dataset contained only financial-related applications.

To achieve this, we first scrape app IDs from multiple pages within each category URL to build a comprehensive dataset [14]. For each app, we collect metadata, including the app name, description, overall rating, total ratings, and a sample

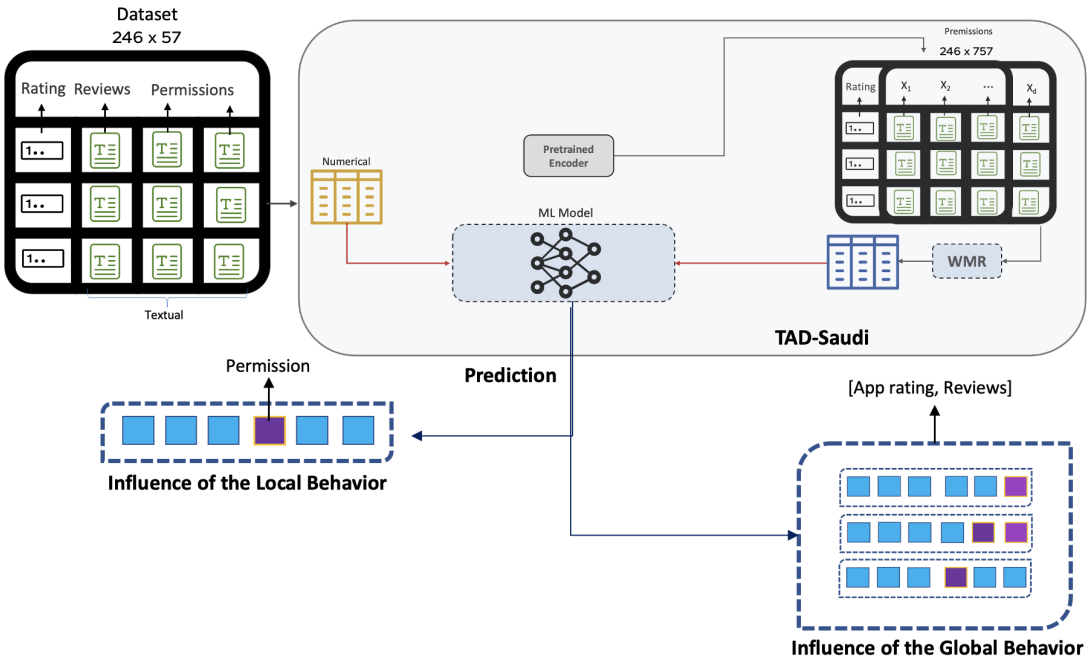


Fig. 1. Our proposed architecture (TAD-SAUDI).

of user reviews. Specifically, we selected the top 20 reviews for each app, with a cap of 15 apps per category, to maintain focus while ensuring diversity across categories. This approach enables us to capture a representative sample of user feedback. The resulting dataset serves as the foundation for analyzing the correlation between app permissions, user perceptions (derived from reviews and ratings), and overall trustworthiness.

In our framework, we operate within a supervised setting, where labeling is a critical step in the development of such an ML model. Specifically, our labeling process involves extracting the top 20 reviews for each application. If at least one of these reviews expresses a significant lack of trust in the application, it is classified as "untrustworthy". Conversely, if none of the reviews raise concerns about trust, the application is labeled as "trustworthy". For each application, the top 20 user reviews are manually examined to identify trust-related content. Reviews explicitly mentioning privacy violations, security concerns, fraudulent activity, data misuse, or suspicious behavior are classified as indicators of untrustworthiness. Reviews focusing on usability, performance issues, software bugs, interface design, or general instability are excluded, as these aspects do not directly reflect users' trust perceptions. Notably, the trust labels in this dataset represent users' subjective perceptions of trustworthiness, rather than verified cybersecurity incidents or objectively confirmed malicious behavior.

The labeling process is conducted manually by the authors. A primary annotator performs the initial labeling, which is subsequently verified by a secondary reviewer to ensure consistency and reliability. As native Arabic speakers, the authors are able to accurately interpret user reviews, particularly since many trust-related concerns are expressed using culturally specific phrases and informal dialects. During the data cleaning process, all non-Arabic reviews, emojis, and irrelevant symbols are removed to avoid potential noise in model performance. This step to ensure that only semantically relevant Arabic content is retained for analysis.

Our constructed dataset consists of a total of 246 applications, with a particular emphasis on financial applications. These applications are categorized into 172 trustworthy apps and 74 untrustworthy apps. Additionally, the dataset includes 3,430 Arabic-language reviews, an equal number of ratings, and 1,581 permissions requested by the apps. The summary of the dataset is illustrated in Table I.

TABLE I. SUMMARY OF THE DATASET

Metric	Value
Total Apps	246
Trustworthy Apps	172
Untrustworthy Apps	74
Total Ratings	3,430
Total Reviews	3,430
Total Permissions	1,581
Total App ID	246
Total App Descriptions	246

As part of our data preparation process, we conducted a human-subjective study involving 51 participants to explore user priorities in selecting trust factors prior to downloading mobile applications. Informed consent was obtained in writing from all participants through a signed consent form before participation.

B. TAD Framework

The goal of TAD framework (Fig. 1) is balance the richness of deep learning with the explainability of simpler models to predict and classify trustworthy apps. Our proposed framework

TAD consists of three main components: 1) A Pretrained Encoder to learn textual representations and extract meaningful and compact feature representations from high-dimensional textual data, 2) Weighted Mean Representation (WMR) to reduce redundancy while preserving essential semantic information, ensuring that the latent representation captures the most discriminative features. Also, we employ attention scores to assign different levels of importance to each feature dimension, and 3) Machine learning model that aims to develop a robust predictor for evaluating the trustworthiness of financial mobile applications, specifically determining whether a user is likely to accept or reject the app. To achieve this, we leverage well-known machine learning classification techniques: XGBoost (XGB) [15], Decision Tree [16], K-Nearest Neighbors (k-NN) [17], Random Forest (RF) [18], and Support Vector Machine (SVM) [19], all of which are well-established in classification tasks. To ensure a comprehensive and fair comparison, we also incorporate a deep learning-based approach, TABNET [20], which is specifically designed for handling tabular data.

Our dataset contains multiple reviews, where each review can have several comments. This structure results in a combination of textual data and straightforward numerical features such as ratings. A naive way to handle such a dataset is to process numerical attributes directly while encoding textual data using methods like label encoding, which assigns a unique integer to each word or phrase. Mathematically, this can be represented as a function mapping words to indices:

$$f(w_i) = i, \quad \forall w_i \in W \tag{1}$$

where, W is the vocabulary and w_i represents each unique word. However, label encoding introduces an inherent ordinality to categorical values, implying an artificial ordering among words (e.g., "good" assigned 1, "bad" assigned 2), which lacks semantic meaning. This causes misleading numerical relationships in machine learning models, as they might incorrectly assume a linear relationship between words based on their assigned numbers. Thus, the text presents additional challenges due to its high-dimensional nature and contextual dependencies.

An alternative solution is to use Deep Neural Networks (DNNs) that employ word embeddings to learn semantic relationships in a high-dimensional space. Mathematically, an embedding function E maps words to dense vectors:

$$E : W \to \mathbb{R}^D, \quad D \gg 1 \tag{2}$$

where, D is the embedding dimension, capturing semantic similarities between words. This allows words with similar meanings to be closer in the vector space, mitigating the limitations of label encoding.

However, as our primary goal is to leverage XAI to understand how user behaviors influence app interactions, relying solely on DNNs [21] can obscure the interpretability of individual financial behaviors (black box models), such as the effect of specific permission requests.

To address both interpretability and effectiveness, we propose an embedding-based feature engineering approach that balances the richness of deep learning with the explainability of simpler models. Instead of directly applying DNNs, we extract pre-trained embeddings for key textual fields, such as description, reviews, and permissions, then reduce them in a structured manner.

Each text field is converted into a high-dimensional vector representation using a pre-trained embedding model. If an embedding of size D is used, then:

$$X_{\text{embedding}} = \{X_1, X_2, ..., X_m\} \tag{3}$$

where, m is the number of textual fields (e.g., app description, reviews, and permissions), and $X_k \in \mathbb{R}^D$ represents the embedding vector of the $k-th$ text field for a given sample. In other words, each text field contributes multiple embedding dimensions—occupying several columns in the tabular dataset.

To make the model interpretable and computationally efficient, we propose the WMR as it is shown in Fig. 1. Mathematically, given

$$x_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]$$

where, x_i represents the feature vector (embedding) for sample i, and D is the embedding dimension. Similarly, the attention weights for sample i are given by:

$$w_i = [w_{i1}, w_{i2}, \ldots, w_{iD}]$$

The weighted mean representation is computed as:

$$\tilde{x}_i = \frac{\sum_{j=1}^{D} w_{ij} x_{ij}}{\sum_{j=1}^{D} w_{ij}} \tag{4}$$

where:

- x_{ij} is the j-th feature of sample i,
- w_{ij} is the j-th attention weight for sample i. The denominator ensures that the weights sum to 1, preserving scale consistency.

This transformation converts each high-dimensional embedding into a single scalar value, significantly reducing complexity while preserving essential semantic information. Without this reduction, the model would struggle to integrate embedding-based features with standard scalar or categorical attributes.

Unlike traditional pooling mechanisms such as attention-based aggregation, which typically assign weights across tokens within a sequence to generate a contextualized sentence embedding, the proposed WMR operates at the feature-dimension level. In other words, instead of weighting tokens, WMR assigns importance to individual embedding dimensions.

The reduced embeddings are concatenated with the remaining non-embedding features. The final feature matrix is denoted as:

$$X' = [X_{\text{structured}}, X_{\text{final}}] \tag{5}$$

where, $X_{\text{structured}}$ consists of traditional tabular features such as categorical metadata and numerical attributes. To mitigate class imbalances, we also apply SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples for underrepresented classes:

$$X_{\text{balanced}}, Y_{\text{balanced}} = \text{SMOTE}(X', Y) \tag{6}$$

IV. Experimental Analysis

We begin this section by providing details on the dataset (Section IV-A) and experimental setup (Section IV-B). Our objective is to first evaluate the performance of our proposed model in terms of predictive accuracy (Section IV-C). Following this, we conduct in-depth analysis (Section IV-D) using SHAP [22] and Section IV-E using LIME [23]) on the top-performing model to investigate: 1) the key features influencing an application's trustworthiness and 2) the relationship between specific features and the acceptance or rejection of an application. Finally, in Section IV-F, we shift our focus to a human-subjective experiment on user trust, aiming to validate our findings through direct user evaluation.

A. Dataset

We compile our dataset through a combination of web scraping techniques and API calls to gather information on finance-related mobile apps from the Google Play Store. The data preparation process is detailed in Section III-A.

The constructed dataset consists of 246 finance-related mobile apps. Given that our goal is to investigate whether users trust mobile applications despite the permissions they request, and to identify the factors influencing their decision-making process, it is crucial to examine the permissions collected for each app. The dataset includes 14 distinct permission categories. Statistics on the permissions associated with the dataset are presented in Table II.

TABLE II. Statistics on Dataset Permissions

Permission	% of Apps
Photos/Media/Files	85
Storage	85
Camera	79
Wi-Fi connection information	77
Location	69
Contacts	54
Phone	53
Device ID and call information	52
Microphone	44
Identity	21
Calendar	9
SMS	7
Device and app history	6.5
Wearable/Activity Sensor Data	0.4

To ensure broad coverage of the Saudi mobile finance ecosystem, the dataset spans 23 finance-related categories on the Google Play Store, including banking, investment, insurance, lending, expense tracking, and financial planning. Within each category, the top-ranked applications were selected to reflect the apps most likely to be encountered by typical users, since marketplace ranking strongly influences user exposure. The resulting class proportions are reported with 95% Wilson confidence intervals: trustworthy applications constitute 69.9% of the dataset (95% CI [63.9%, 75.3%]) and untrustworthy applications 30.1% (95% CI [24.7%, 36.1%]). A power analysis confirms that at $n = 246$, $\alpha = 0.05$, and $power = 0.80$, the dataset is adequate to detect medium effect sizes (Cohen's $d \geq 0.389$) between the two classes.

Two sources of potential bias are acknowledged. First, the dataset is restricted to applications available on the Google Play Store and does not include Apple App Store apps or applications distributed through alternative Arabic app stores; findings may therefore not transfer directly to iOS users. Second, the top 20 reviews per app sampling strategy favors reviews surfaced by Google's ranking algorithm, which tends to prioritize reviews marked as helpful or those that are more recent, potentially underrepresenting older or less-visible user concerns. We mitigate these effects by drawing from 23 distinct finance categories rather than concentrating on a single segment, which broadens topical and demographic coverage.

The dataset contains 172 trustworthy and 74 untrustworthy applications, yielding an approximate 2.32:1 ratio. This imbalance reflects the natural distribution of trust perceptions in the financial app marketplace, where most listed applications belong to established institutions and receive predominantly positive trust signals. To prevent the majority class from masking poor minority-class performance, we apply stratified k-fold cross-validation, which preserves the class ratio in each fold, and report macro-averaged precision, recall, and F1-score in addition to overall accuracy.

B. Experimental Settings

We randomly select 75% of the dataset for training and 25% for testing. For the Logistic Regression model, the maximum number of iterations is set to 500. For k-NN, we choose 5 as the number of nearest neighbors. Logarithmic loss is used as the evaluation metric to track the performance of the XGB model during training.

To effectively represent the semantic and contextual meaning of Arabic text, we employ the AraBERT pre-trained embedding model [24], which generates dense vector representations of the reviews. These embeddings serve as the textual input features for all models, enabling a deeper understanding of linguistic nuances and improving the overall accuracy of trustworthiness classification. To evaluate the performance of our proposed framework, TAD-Saudi, we use the following metrics:

- Accuracy: Measures the proportion of correct predictions out of the total number of predictions.
- Precision: Represents the ratio of correctly predicted positive instances among all instances predicted as positive.
- Recall: Indicates the ratio of correctly predicted positive instances relative to the total number of actual positive instances.

- F1 Score: The mean of precision and recall, providing a balanced evaluation of the model's performance.

The foundation of our TAD-Saudi framework lies in the proposed encoding schema. We compare our TAD-Saudi framework with three common encoding techniques: Label Encoding [25], One-Hot Encoding[26], and Binary Encoding[27]. To ensure a comprehensive comparison, we apply each of these encoding methods, along with our proposed framework, to the following models: RF [18], Gradient Boosting [28], XGB [15], TABNET [20], Logistic Regression [29], K-NN [17], SVM [30], and Naive Bayes [31]. A brief summary of these models is as follows:

- RF: A model that combines multiple decision trees using randomly sampled data and features.
- Gradient Boosting: An optimization-based ensemble method that builds additive models, often using decision trees, and minimizes loss via gradient descent.
- XGB: A scalable tree boosting system designed for high-performance machine learning on large datasets.
- TABNET: An interpretable deep learning architecture for tabular data that uses sequential attention to efficiently select important features.
- Logistic Regression: A statistical method used to model the relationship between a binary dependent variable and one or more independent variables.
- K-NN: A supervised machine learning algorithm that classifies data points based on the majority class among their 'k' closest neighbors in the feature space.
- SVM: One of the earlier supervised machine learning algorithms that has been instrumental in both classification and regression tasks.
- Naive Bayes: A simple probabilistic classifier based on Bayes' theorem, assuming that features are independent given the class label.

C. Performance Evaluation

The performance evaluation results presented in Table III highlight the effectiveness of the proposed TAD-Saudi framework in comparison to conventional encoding techniques, namely Label Encoding, One-Hot Encoding, and Binary Encoding. The findings demonstrate that TAD-Saudi consistently outperforms all traditional encoding methods across multiple machine learning models, achieving higher accuracy, precision, recall, and F1 scores.

Among the traditional encoding methods, Label Encoding, One-Hot Encoding, and Binary Encoding yield comparable performance, with accuracy scores ranging primarily between 0.60 and 0.71. The best-performing models under these methods include Logistic Regression (0.74, Label Encoding), Gradient Boosting (0.71, Label Encoding), and XGBoost (0.69, One-Hot & Binary Encoding). However, these models fail to exceed the 0.74 accuracy threshold, indicating the limitations of conventional encoding approaches in capturing the complexities of the dataset.

In contrast, TAD-Saudi achieves a significant performance boost, with XGBoost and Random Forest each attaining the highest accuracy of 0.85, followed by Gradient Boosting and Naïve Bayes at 0.81. These models also exhibit higher precision, recall, and F1 scores, reinforcing the dominance of TAD-Saudi in learning robust representations. The observed improvements across different models indicate that the TAD-Saudi framework enhances feature encoding, leading to better generalization and decision-making capabilities.

A key observation is the poor performance of TABNET across all settings, achieving accuracy values below 0.50 in most cases. As a result, deep learning-based methods may require additional tuning or architectural modifications to effectively handle the tabular data in the given dataset. Furthermore, the nature of tabular data, with its high sparse relationships, could pose significant challenges for models designed primarily for structured inputs like images or sequences [2]

Overall, these results validate the effectiveness of TAD-Saudi in improving predictive accuracy, particularly in handling categorical data more efficiently than traditional encoding techniques. The model demonstrates a clear advantage in capturing complex structured relationships within the data, leading to enhanced performance. This indicates that TAD-Saudi's approach to feature representation and processing offers a promising alternative for models dealing with large, structured datasets, where traditional methods often struggle.

D. Global Analysis

To understand the global behavior and underlying logic behind trustworthy and untrustworthy apps, we utilize SHAP [22], a unified method that offers interpretable explanations for the predictions made by machine learning models. By analyzing the feature contributions for apps, we can gain insights into which factors influence the classification of an app as trustworthy or untrustworthy. This allows us to identify patterns and relationships that clarify which factors are most critical in the determination of app trustworthiness, shedding light on the key characteristics that distinguish these two categories. We select the best-performing model, XGBoost, in our TAD-Saudi framework for our analysis to provide a reliable and accurate explanation of its behavior.

Fig. 2 presents the SHAP analysis of the top 20 factors that impact the trustworthiness of an app on our proposed framework, where the features are listed according to their total contribution (sum of SHAP values) to the final prediction, with the top feature being the most important. Each point represents an app in the study. The position of each point on the x-axis indicates the impact that feature has on the classifier's prediction for a given app, with the color gradient (ranging from blue to red) reflecting the feature values from low to high. On the x-axis, positive SHAP values indicate the trustworthiness of the app, whereas negative SHAP values reflect untrustworthiness. This visualization allows us to observe the relative influence of each feature on the prediction, providing a clear understanding of which factors play a dominant role in shaping an app's classification as trustworthy or untrustworthy.

The interpretation in Fig. 2 shows several key observations. First, permissions play a critical role in determining trust-

TABLE III. PERFORMANCE EVALUATION

Model	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
	Label Encoding				One-Hot encoding			
RF [18]	0.65	0.62	0.65	0.63	0.65	0.61	0.65	0.62
Gradient Boosting [28]	0.71	0.70	0.71	0.70	0.63	0.61	0.63	0.62
XGB [15]	0.68	0.67	0.68	0.67	0.69	0.67	0.69	0.67
TABNET [20]	0.48	0.55	0.48	0.50	0.48	0.55	0.48	0.50
Logistic Regression[29]	0.74	0.74	0.74	0.70	0.73	0.71	0.73	0.69
K-NN [17]	0.63	0.59	0.63	0.60	0.63	0.59	0.63	0.60
SVM [30]	0.69	0.48	0.69	0.57	0.69	0.48	0.69	0.57
Naive Bayes [31]	0.66	0.76	0.66	0.67	0.66	0.76	0.66	0.67
	Binary encoding				TAD-Saudi			
RF [18]	0.65	0.61	0.65	0.62	0.85	0.86	0.85	0.85
Gradient Boosting [28]	0.63	0.61	0.63	0.62	0.81	0.81	0.81	0.81
XGB [15]	0.69	0.67	0.69	0.67	0.85	0.85	0.85	0.85
TABNET [20]	0.48	0.55	0.48	0.50	0.43	0.43	0.43	0.43
Logistic Regression[29]	0.73	0.71	0.73	0.69	0.80	0.82	0.80	0.80
K-NN [17]	0.63	0.59	0.63	0.60	0.67	0.72	0.67	0.66
SVM [30]	0.69	0.48	0.69	0.57	0.80	0.81	0.80	0.80
Naive Bayes [31]	0.66	0.76	0.66	0.67	0.81	0.86	0.81	0.81

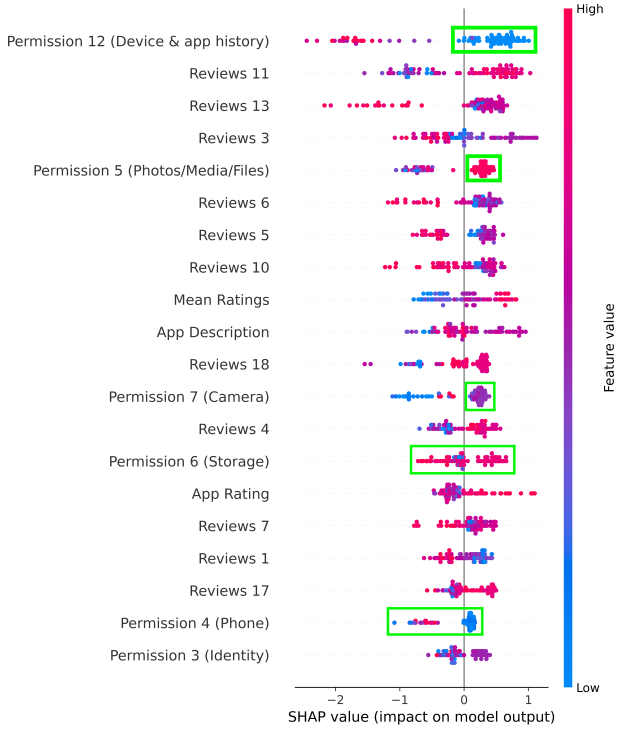


Fig. 2. SHAP analysis of the top 20 factors influencing app trustworthiness in the proposed framework. Features are ranked by their total SHAP contribution, with higher values indicating greater importance. Each point represents an app, where color denotes feature value (blue = low, red = high) and the x-axis shows impact direction—positive values indicate trustworthiness, while negative values reflect untrustworthiness.

worthiness, with Device & App History, Photos/Media/Files, Camera, Storage, Phone, and Identity permissions appearing among the most influential features. Apps requesting extensive permissions, particularly those related to sensitive data access, tend to have a negative SHAP impact (untrustworthy apps), indicating that such permissions are more commonly associated with untrustworthy apps. This aligns with security concerns where excessive permissions may indicate potential privacy risks [32].

While permissions generally exhibit a negative correlation with app trustworthiness in the SHAP summary plot, certain permissions demonstrate mixed or even positive contributions, indicating that their impact is context-dependent. For instance, Permission 12 (Device & App History) displays SHAP values that extend into the positive range, indicating that some apps requiring this permission—such as system utilities or security applications—are still classified as trustworthy. Similarly, Permission 5 (Photos/Media/Files) and Permission 7 (Camera) show instances where high feature values contribute positively to trustworthiness, likely reflecting legitimate use cases in photo-editing, document scanning, and video conferencing apps. Additionally, Permission 6 (Storage) and Permission 4 (Phone) exhibit mixed SHAP contributions, implying that while they are often associated with untrustworthy apps, certain well-rated applications that transparently justify these permissions remain trustworthy.

These observations explain that ratings and reviews alone are not sufficient to determine an app's trustworthiness, as permissions also play a crucial role in the classification process. While highly rated and frequently reviewed apps are often perceived as more reliable [33], the SHAP analysis indicates that certain permissions can significantly impact trustworthiness negatively. This highlights the importance of transparency and responsible permission management, reinforcing that trustworthiness is not shaped only by user ratings and reviews but by a combination of user feedback, feature permissions, and the app's overall reputation. While high ratings and positive reviews contribute to an app's credibility, they do not fully capture the potential risks associated with excessive or sensitive permission requests.

E. Local Analysis

To complement our global analysis of app trustworthiness using SHAP, we perform a nuanced local interpretability analysis using LIME [23]. LIME produces instance-specific explanations, enabling us to uncover how specific combinations of features influence the classification of apps as trustworthy or untrustworthy. This refined perspective exposes decision-making patterns that may remain hidden in global summaries, offering deeper insights into user-perceived and model-learned trust signals.

Our analysis focuses on the XGBoost model, chosen for its robust predictive performance in the TAD-Saudi framework. LIME is applied to each sample in the test set to extract a feature attribution vector, where each value quantifies the positive or negative contribution of a feature to the model's decision. Positive values support a classification of trustworthy, while negative values indicate influence toward untrustworthy. To identify dominant patterns of local behavior, we clustered these vectors using KMeans (k=5) [34].

Fig. 4 presents a heatmap showing the average LIME feature attributions across the five clusters. Each row corresponds to a cluster, and each column to a feature. Red cells indicate features that consistently support trust predictions, while blue cells indicate features that consistently detract from trust.

Our cluster-wise LIME analysis reveals how variations in user behavior and perception drive differences in trust formation across app groups. In Cluster 0, trust predictions are positively influenced by early reviews such as review_5 and review_6, along with permission_7 (Camera). However, later reviews like review_19 and review_20 show strong negative impacts. This shift indicates that users initially responded positively but later became dissatisfied—potentially due to app updates, intrusive notifications, or emerging performance issues. The fact that the Camera permission maintains a positive association with trust indicates contextual acceptance behavior, where users justify certain permissions when they align with the app's functionality (e.g., photo-sharing or video features). This pattern aligns with trust heuristics in digital behavior, where familiarity and early positive experience temporarily override privacy concerns.

In Cluster 1, user behavior reflects a contrasting cognitive pattern. Trust is supported by positive review features (review_8, review_17, review_19), but sharply decreases in the presence of permission_3 (Identity) and permission_10 (Device ID and Call Info). This reveals a privacy-centered behavioral profile in which users exhibit a greater sensitivity to permissions involving personal or device identifiers. Even when social feedback (reviews) is favorable, these users value privacy safeguards more than social validation, indicating a risk-averse trust model. Overall, the cluster patterns highlight that user trust is not purely sentiment-driven but context-dependent, balancing perceived utility, social validation, and privacy risk in complex behavioral trade-offs.

Clusters 2 and 3 exhibit a similar trust logic, where reputational signals—such as rating, review_17 (in Cluster 2), and review_11 (in Cluster 3)—positively influence trust predictions. However, both clusters show strong negative contributions from permission_10 (Device ID and Call Info) and select reviews (review_13 in Cluster 3 and description in Cluster 2). This pattern reveals a user tendency to favor high ratings and positive feedback, even in the presence of potential privacy risks—indicating that strong reputational signals often take precedence over deeper concerns about app behavior. Finally, Cluster 4 reflects a context-aware decision pattern. Positive signals such as review_20 and review_12 increase trust, while review_11 and permission_10 decrease it. This balance indicates that users integrate both reputational and risk considerations, granting trust only when favorable cues substantially exceed perceived privacy risks—demonstrating a more deliberate trust formation process.

To evaluate feature reliability, we analyze the standard deviation of LIME attributions for each feature (Fig. 3). Features such as review_11, review_17, and review_19 display high variability, indicating that their influence on model predictions is context-dependent and unstable, varying considerably across different instances. Conversely, features like permission_3 (Identity) and permission_12 (Device & App History) exhibit low variability with consistently negative attributions, indicating they are reliable and stable indicators of untrustworthiness.

Overall, while SHAP reveals overarching trends in feature importance, LIME uncovers the diversity in localized decision-making logic. A key insight is that users often favor apps with high ratings and early positive reviews, even when such apps request permissions typically associated with untrustworthiness (e.g., Camera, Identity). This suggests that visible reputational cues can obscure privacy concerns, leading to over-trust in potentially risky apps. Moreover, the influence of permission-related features varies by context. For example, permission_12 might be accepted for security tools but viewed suspiciously in general-purpose apps.

F. A Human-Subjective Experiment on User Trust

As our findings indicate, Section IV-D, permissions play a critical role in shaping app trustworthiness. While high ratings and positive reviews are commonly seen as indicators of reliability, they alone do not provide a complete picture of an app's trustworthiness. To gain a deeper understanding of this from a human perspective, we conduct a human-subjective test in which we survey 51 users to investigate their priorities when selecting trust factors before downloading apps. All the participants ranged in age from 20 to 50 years and had diverse academic and professional backgrounds. They are Saudi Arabia residents, native Arabic speakers, and regular mobile

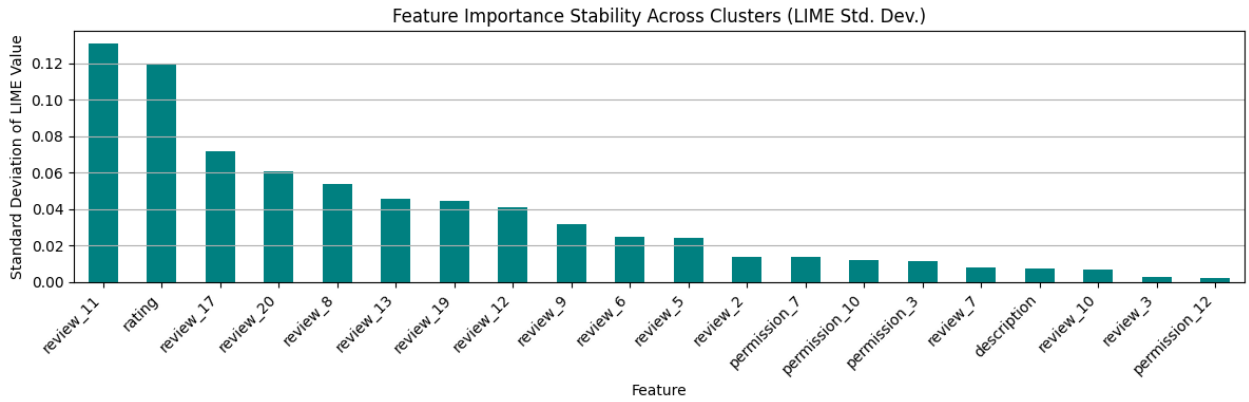


Fig. 3. Feature importance stability across clusters.

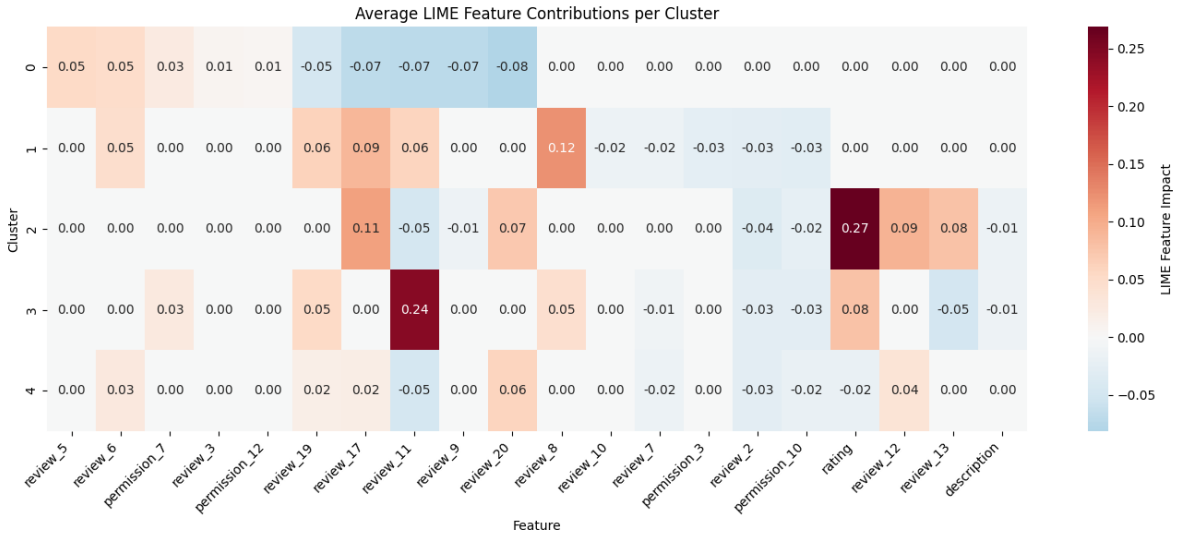


Fig. 4. Average LIME feature contributions per cluster. The heatmap shows the average LIME feature attributions across five clusters. Each row represents a cluster, and each column a feature. Red cells indicate features that positively contribute to trust predictions, while blue cells denote features that negatively influence trust.

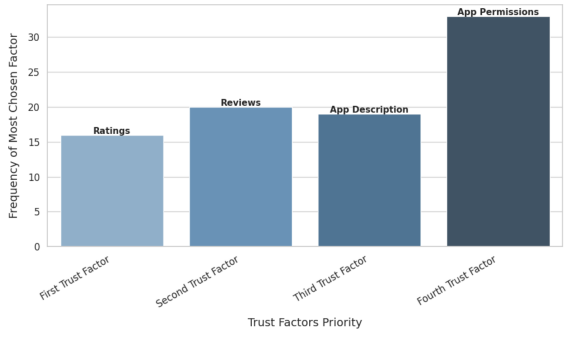


Fig. 5. User Priorities in App Trust: Order of Trust Factors (First to Fourth), where the sequence of trust factors, indicating that the first trust factor is the most important, followed by the others in order of importance.

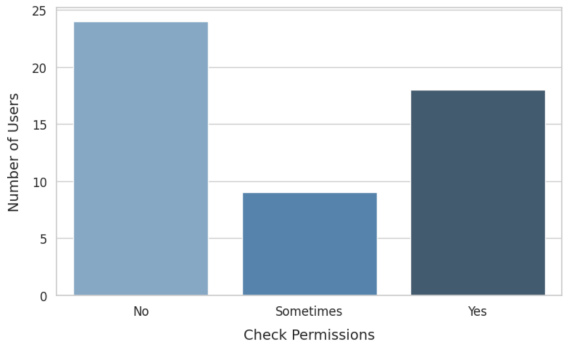


Fig. 6. User responses to checking app permissions before downloading.

device users, either iPhone or Android. They are general users with varying levels of digital experience, which allowed us to gather feedback that reflected real-life experiences.

The results, as shown in Fig. 5, reveal that app permissions emerged not as the most important factor, with 35 users selecting it as their fourth priority (after factors like ratings, reviews, and app descriptions). This finding indicates that while app

permissions are still a significant factor for app trustworthiness, they are not the first factor users prioritize when downloading apps. This indicates that a broader awareness campaign is needed to emphasize the importance of permissions, highlighting their role in ensuring user security and privacy. A more comprehensive understanding of permission management could potentially shift users' priorities, making it a more prominent factor in their trust evaluation process.

Further analysis of the question "Do Users Check App Permissions Before Download?" (Fig. 6) reveals a divided response of the 51 participants, 24 users indicate that they do not check app permissions, while 9 report occasionally checking, and 18 state that they always check permissions. This reveals that a significant portion of users does not consistently check app permissions, which indicates a lack of awareness or concern regarding potential privacy and security risks associated with app permissions.

Overall, in alignment with recent SHAP analysis, our findings indicate that certain app permissions can negatively impact trust, even if an app boasts high ratings or favorable reviews. Therefore, transparent permission management and clear communication regarding the purpose of these permissions are vital for fostering user trust. This combination of user feedback, careful permission handling, and an app's overall reputation forms a more comprehensive and accurate understanding of its trustworthiness, highlighting that trust is shaped by a balance of factors, including app permissions.

V. Limitations

While the constructed dataset of 246 finance-related mobile applications provided valuable insights into user trust and app permissions, its relatively small size presents a potential limitation for machine learning generalization. The modest number of samples may restrict the model's ability to capture broader behavioral variability across different app categories and user contexts.

The findings of this study apply most directly to Arabic-language financial applications used in the Saudi context. Generalization to other domains (such as non-financial app categories), other languages, or other regional markets, including other Gulf or broader Arabic-speaking populations with different regulatory frameworks and cultural contexts, would require further validation. The methodological framework itself, however, combining permission analysis, review embeddings, and explainable AI techniques, is transferable, and we view replication across additional markets and domains as an important direction for future work.

VI. Conclusion

The proposed TAD-Saudi framework provides a significant step forward in the evaluation of mobile app trustworthiness by bridging interpretability and performance in machine learning–based detection. Beyond its technical contribution, the study offers practical implications for multiple stakeholders. For app developers, TAD-Saudi serves as an interpretable benchmark to identify risky permission patterns and improve transparency in app design. For regulators and cybersecurity authorities, it enables data-driven policy formulation and supports more reliable certification processes for financial and e-service applications. For end-users, it enhances awareness of privacy risks and promotes trust in secure digital ecosystems aligned with Saudi Vision 2030.

Future research will expand the dataset to include a larger and more diverse range of applications across different app categories and regions. Further investigations could explore cross-country validation to assess generalizability in other Gulf and international contexts, as well as integration with real-time app store monitoring systems to enable dynamic trust evaluation. These directions aim to strengthen the scalability, robustness, and real-world applicability of the TAD-Saudi framework.

Author Contributions

Raed was responsible for the methodology, analysis, writing, and supervision of the work. Maryam contributed to data curation and editing parts of the manuscript. All authors reviewed the manuscript.

Funding

This research is supported by a grant (No. $CRPG-25-3082$) under the Cybersecurity Research and Innovation Pioneer grant, provided by the National Cybersecurity Authority (NCA) in the Kingdom of Saudi Arabia.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

The data is available at the following link: https://github.com/maryamghamdi/TAD_Saudi

Ethical Approval

This study has received ethical approval from the Saudi Electronic University under approval number SEUREC-4648. All procedures involving human participants were conducted in accordance with the ethical standards of the institutional research committee and with the Helsinki Declaration.

References

[1] M. Alrizq and A. Alghamdi, "Customer satisfaction analysis with saudi arabia mobile banking apps: a hybrid approach using text mining and predictive learning techniques," *Neural Computing and Applications*, vol. 36, no. 11, pp. 6005–6023, 2024.

[2] R. Alharbi, S. Chan-Olmsted, H. Chen, and M. T. Thai, "Deep learning framework with multi-perspective social behaviors for vaccine hesitation," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 140, 2024.

[3] S. M. Habib, N. Alexopoulos, M. M. Islam, J. Heider, S. Marsh, and M. Müehlhäeuser, "Trust4app: automating trustworthiness assessment of mobile applications," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 124–135.

[4] M. Kuehnhausen and V. S. Frost, "Trusting smartphone apps? to install or not to install, that is the question," in *2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 2013, pp. 30–37.

[5] O. Akgul, S. T. Peddinti, N. Taft, M. L. Mazurek, H. Harkous, A. Srivastava, and B. Seguin, "A decade of privacy-relevant android app reviews: large scale trends," in *Proceedings of the 33rd USENIX Conference on Security Symposium*, ser. SEC '24. USA: USENIX Association, 2024.

[6] P. Nema, P. Anthonysamy, N. Taft, and S. T. Peddinti, "Analyzing user perspectives on mobile app privacy at scale," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022, pp. 112–124.

[7] Ö. Turgut, M. E. Ayhan, Ö. Coşkun, İ. Kök, and S. Özdemir, "Trustworthy and interpretable machine learning models for financial fraud detection," in *2025 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE, 2025, pp. 1–6.

[8] Q. Xu, Y. Liao, Q. Li, J. Zhang, Z. Song, L. Wang, and X. Yuan, "Shap-based interpretable models for credit default assessment using machine learning," in *2024 14th International Conference on Software Technology and Engineering (ICSTE)*. IEEE, 2024, pp. 213–217.

[9] H. Harkous, S. T. Peddinti, R. Khandelwal, A. Srivastava, and N. Taft, "Hark: A deep learning system for navigating privacy feedback at scale," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 2469–2486.

[10] M. Hatamian, J. Serna, and K. Rannenberg, "Revealing the unrevealed: Mining smartphone users privacy perception on app markets," *Computers & Security*, vol. 83, pp. 332–353, 2019.

[11] T. Rani, S. S. Sakthy, P. Kalaichelvi, P. S, and V. S, "Fake app detection using sentiment analysis," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, 2023, pp. 1–6.

[12] H. Zaki, M. Saad, and M. R. Rasheed, "A robust machine learning framework for fraudulent mobile app detection," *VFAST Transactions on Software Engineering*, vol. 12, no. 4, pp. 27–36, 2024.

[13] Google, "Google play store," 2025, accessed: 2025-01-07. [Online]. Available: https://play.google.com/

[14] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application." *International Journal of Advances in Soft Computing & Its Applications*, vol. 13, no. 3, 2021.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[16] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[17] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, 2009.

[18] L. Breiman, "Random forests," *UC Berkeley TR567*, 1999.

[19] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *CVPR*, 2006.

[20] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[21] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, vol. abs/2006.11371, 2020.

[22] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[24] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 9.

[25] D. Shah, Z. Y. Xue, and T. M. Aamodt, "Label encoding for regression networks," *arXiv preprint arXiv:2212.01927*, 2022.

[26] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[27] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," *Advances in neural information processing systems*, vol. 19, 2006.

[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[29] J. Berkson, "Application of the logistic function to bio-assay," *Journal of the American statistical association*, vol. 39, no. 227, pp. 357–365, 1944.

[30] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[31] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, pp. 103–130, 1997.

[32] K. Degirmenci, "Mobile users' information privacy concerns and the role of app permission requests," *International Journal of Information Management*, vol. 50, pp. 261–272, 2020.

[33] H. Sällberg, S. Wang, and E. Numminen, "The combinatory role of online ratings and reviews in mobile app downloads: an empirical investigation of gaming and productivity apps from their initial app store launch," *Journal of Marketing Analytics*, vol. 11, no. 3, pp. 426–442, 2023.

[34] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5. University of California press, 1967, pp. 281–298.

Benchmarking Deep Learning for $PM_{2.5}$ Forecasting in IoT-Based Smart Cities: GCN Spatial Encoding and Transformer Temporal Modeling

Abdessamad BADOUCH[1], Kaoutar BELHOUCINE[2]
Faculty of Sciences, Ibn Zohr University, Agadir, Morocco[1]
Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco[2]

***Abstract*—Graph convolutional networks are widely used in air quality forecasting, yet their benefit over simpler approaches remains insufficiently validated. This study presents a fully reproducible benchmark comparing five model families (ARIMA, XGBoost, LSTM, CNN-LSTM, and a GCN+Transformer model) on the public Beijing Multi-site Air Quality Dataset. All experiments share identical preprocessing, three random seeds (confidence intervals reported as exploratory), and a chronological split, with full code availability. ARIMA outperforms deep learning on aggregated data, due to strong temporal autocorrelation. On individual stations, Transformer-based models achieve the best performance through improved temporal modeling. A systematic multi-node analysis reveals that graph-based spatial aggregation can degrade performance under highly homogeneous conditions: when monitoring stations exhibit strong correlations, a simple linear baseline outperforms the GCN topologies tested here. Alternative spatial encoders (e.g., GAT, learnable adjacency) may behave differently and remain an open question. These findings define a practical regime in which GCN-based spatial aggregation provides no benefit over linear baselines. A preliminary compression experiment on Apple M1 reports a 42% inference-latency reduction via 8-bit quantization; this is indicative only, and validation on Raspberry Pi 4 or NVIDIA Jetson Nano is identified as required follow-up.**

***Keywords*—Air quality forecasting; graph neural networks; IoT smart cities; transformer; $PM_{2.5}$ prediction**

I. Introduction

Urban air pollution is one of the most serious environmental health challenges of our time. $PM_{2.5}$ particles, with aerodynamic diameter below 2.5 micrometres, penetrate deep into lung tissue and are associated with cardiovascular and respiratory disease, premature mortality, and impaired cognitive development [2], [3]. In cities across China, India, and the Middle East, daily $PM_{2.5}$ concentrations regularly exceed WHO guidelines by factors of five to ten. Forecasting these concentrations accurately, and doing so in real time, is a prerequisite for any early warning or emergency response system.

Smart city infrastructure changes what is technically possible here. Networks of low-cost IoT sensors now generate continuous, high-resolution spatiotemporal streams from dozens of monitoring stations within a single city. This spatial dimension is what makes graph neural networks attractive: if stations at different locations carry different information, encoding their spatial relationships should help the model. Whether inter-station diversity is high enough for graph aggregation to help, rather than simply average away useful local variation, depends on the specific city and meteorological regime. That question motivates this study.

This study does not propose a new architecture. It provides a reproducible benchmark on public data, with a clear empirical result: on Beijing's monitoring network, graph spatial encoding does not improve forecasting performance over a linear baseline. Published claims about GCN superiority in air quality forecasting consistently appear on proprietary, spatially heterogeneous datasets [20], [22], often multi-city or multi-provincial, without testing whether the spatial component would remain useful if correlations were high and uniform. Our benchmark fills that gap by providing the baseline case. Negative results on public reproducible benchmarks are an undervalued but essential contribution to the field: without them, the community cannot know which spatial modeling assumptions are safe to import from traffic forecasting, where GCN consistently helps, to air quality forecasting, where the spatial precondition may not hold.

Showing that GCN does not help in one spatially homogeneous setting does not by itself establish exactly when it does help. Testing that requires a second dataset from a heterogeneous network, which we identify as the most important direction for follow-up work. The contribution here is a documented, reproducible reference case, grounded in the meteorology of the Beijing basin.

The contributions of this study are as follows. First, a reproducible benchmark comparing five model families on UCI #501 is provided (anonymized repository; URL withheld during review and to be disclosed upon acceptance), with fixed seeds and public code, addressing the reproducibility gap in prior work [20], [21], [22]. Second, a multi-signal analysis identifies when ARIMA is competitive (aggregated, smooth signals) and when it is not (per-station noisy data), with the Transformer and GCN components disentangled in both settings. Third, a reproducible negative result on GCN spatial encoding is established for networks with near-uniform correlations, with a physical explanation rooted in Beijing's meteorological dynamics. Fourth, a preliminary edge deployment analysis reports a 42% latency reduction via 8-bit quantization on Apple M1 hardware, offered as an initial feasibility indicator for IoT deployment.

II. Related Work

A. Classical and Statistical Methods

ARIMA and its seasonal variant SARIMA remain the standard baselines in time series forecasting. They are well understood, interpretable, and require no training data beyond the target series itself. Their main limitation is the linearity assumption: ARIMA captures autocorrelation structure well but cannot model abrupt nonlinear events such as dust storms or holiday traffic changes. Recent work has also explored statistical process control approaches such as modified EWMA charts for autocorrelated $PM_{2.5}$ monitoring [23], which complement forecasting models for anomaly detection use cases. Beyond $PM_{2.5}$, related assessment frameworks have been applied to other pollutants such as SO_2 [24], confirming the general relevance of statistical baselines in air quality studies. Our benchmark confirms R^2=0.973 for ARIMA on the aggregated signal, consistent with prior comparisons [18]. This result is a product of the aggregation process, as we discuss in Section IV.

XGBoost and gradient boosting methods have also been applied to air quality forecasting with good results on tabular feature sets [1]. Hybrid pipelines combining LSTM, Random Forest, and XGBoost have also been proposed for real-time AQI prediction [25], combining the complementary strengths of recurrent and tree-based models. They handle nonlinearity and work well on moderate-size datasets without tuning recurrent architectures. Their limitation is the same as ARIMA's: they treat each time step independently unless explicit lag features are engineered, and they have no native mechanism for spatial reasoning.

B. Deep Learning Models

LSTM networks [16] were among the first deep learning architectures applied to air quality forecasting, and they remain strong baselines. The gated mechanism handles long-range temporal dependencies better than vanilla RNNs, but training is sequential and parallelism is limited. Several studies have combined convolutional layers with LSTM to capture both local spatial patterns and temporal dynamics [10], [11], [12], with reported improvements over LSTM-only baselines in spatially structured settings.

Transformer architectures [17] arrived later in this application domain. The self-attention mechanism processes all timesteps in parallel, is more efficient on GPU hardware, and handles long-range dependencies without the vanishing gradient issues that affect LSTMs. AirFormer [22] demonstrated the Transformer's capacity to scale to 1,085 stations across China, though on proprietary data and without systematic baselines. Our ablation confirms that the Transformer is the decisive component in our framework, contributing ΔR^2=0.111 relative to an equivalent GCN+LSTM architecture.

C. Graph Neural Networks in Air Quality

GNNs have become popular for air quality forecasting because monitoring stations are naturally modeled as graph nodes. Zhang et al. [10] combined temporal difference features with graph Transformer networks and reported improvements over CNN-LSTM on multi-station data in China. Liang et al. [20] developed a Res-GCN model incorporating satellite imagery on 34 Beijing and 27 Tianjin stations, using a dynamic GCN structure to capture evolving spatial patterns. These results are encouraging but rely on multi-city or multi-provincial settings where inter-station spatial diversity is inherently higher. Ghose et al. [6] applied a graph-like neighbor influence model in an Indian smart city context, showing that nearby station information helps when pollution sources are spatially heterogeneous.

D. Structured Taxonomy: Reproducibility and Ablation Gaps

Table I surveys representative works spanning classical statistical methods to graph-based and Transformer architectures, making clear where this study sits within the literature. Eight evaluation criteria are assessed: inclusion of a classical baseline (ARIMA or similar), machine learning baselines (XGBoost/RF), recurrent DL (LSTM/GRU), CNN-based spatial features, graph neural network component, Transformer component, controlled GCN ablation against a linear baseline, and full reproducibility (public data and code). Symbols: ✓ present; x absent; $\sim$ partial.

Several patterns stand out in Table I. Among surveyed works, none simultaneously provides all eight features. The most common gap is the absence of GCN ablation: GCN components are commonly added without testing whether GCN outperforms a linear baseline on the same data. Similarly, full reproducibility on public data with code remains rare. Among the works using GCN [6], [10], [20] or similar graph structures, none ablate the spatial component against a linear baseline, leaving open the question of whether the reported gains come from the graph aggregation itself or from the other components of the architecture. Table II summarises method-by-method gaps and how this study addresses each of them.

III. Methodology

A. Problem Formulation and Graph Setup

The problem is framed as multivariate time series regression. At each hourly timestep t, a station i reports a vector of six measurements: $PM_{2.5}$ concentration (the target), NO_2, temperature (TEMP), atmospheric pressure (PRES), dew-point temperature (DEWP), and wind speed (WSPM). The model receives a sequence of 24 consecutive hourly observations (window length $T = 24$) and predicts $PM_{2.5}$ at timestep $t+1$.

In the multi-node setting, the 12 stations are represented as graph nodes. Two graph construction strategies are compared. In the *distance-based* topology, edges connect each station to its $k = 5$ nearest neighbors by Haversine distance, with edge weights $w_{ij} = 1/d_{ij}$ inversely proportional to distance. In the *correlation-based* topology, edges connect the $k = 5$ stations with the highest Pearson correlation of $PM_{2.5}$ time series, computed exclusively on the training split to prevent data leakage. In the aggregated benchmark, the 12 station signals are averaged into a single hourly series before model input; in that setting the GCN has only one node and therefore performs a linear projection of the feature vector rather than any neighborhood aggregation.

The choice $k = 5$ was validated by a sensitivity analysis over $k \in \{3, 5, 7, 10\}$: performance under the distance-based topology varies by less than ΔR^2=0.004 across all values

TABLE I. COMPREHENSIVE TAXONOMY OF AIR QUALITY FORECASTING STUDIES (2019–2025). ARIMA = CLASSICAL STATISTICAL BASELINE; ML = TREE/ENSEMBLE ML; LSTM = RECURRENT DL; CNN = CONVOLUTIONAL SPATIAL FEATURES; GCN = GRAPH NEURAL NETWORK; TRANS. = TRANSFORMER; ABLAT. = CONTROLLED GCN VS. LINEAR ABLATION; REPRO. = PUBLIC DATA AND CODE

Study	Dataset	ARIMA	ML	LSTM	CNN	GCN	Trans.	Ablat.	Repro.
Kow et al. [5] 2022	Taiwan	✓	x	x	✓	x	x	x	x
Zhang et al. [10] 2022	Beijing	x	x	✓	x	✓	✓	x	x
Zhang et al. [11] 2022	Chinese cities	x	x	✓	✓	x	x	x	x
Ghose et al. [6] 2022	India, smart city	x	x	✓	x	∼	x	x	x
Krishan et al. [16],[3]	Multi-dataset	✓	✓	✓	x	x	x	x	x
Cican et al. [8] 2023	Bucharest, Romania	✓	✓	✓	x	x	x	x	x
Srivastava [12] 2023	Multi-station	x	x	✓	x	x	x	x	x
Soh et al. [21] 2023	Beijing+Taizhou	✓	x	x	x	x	✓	x	x
Hameed et al. [2] 2023	Doha, Qatar	x	x	✓	✓	x	x	x	x
Ansari & Quaff [13] 2025	Azamgarh, India	✓	✓	✓	x	x	x	x	x
Ansari & Quaff [14] 2025	Azamgarh, India	✓	✓	✓	x	x	x	x	x
Tello-Leal et al. [7] 2024	Mexico, urban	x	✓	✓	x	x	x	x	x
Liang et al. [20] 2024	Beijing+Tianjin	x	x	x	✓	✓	x	x	x
Rahmani et al. [9] 2024	France, multi-city	x	x	✓	x	x	x	x	x
Shahbazi et al. [1] 2024	Multi-region	✓	✓	✓	x	x	x	x	x
Alsabagh et al. [19] 2025	Beijing UCI #501	✓	x	✓	✓	x	x	x	x
This work (2025)	**Beijing, UCI #501**	✓	✓	✓	✓	✓	✓	✓	✓

TABLE II. COMPARISON OF RELATED WORK: GAPS ADDRESSED BY THIS STUDY

Study	Dataset	Method	Key Limitation	Addressed
Du et al. [4]	Beijing UCI	CNN-LSTM	No Transformer; no GCN ablation; no edge deployment	✓
Kow et al. [5]	Taiwan	Deep CNN	Image input only; single modality	✓
Hameed et al. [2]	Doha	Multimodal DL	Proprietary data; no GNN; no edge	✓
Zhang et al. [10]	Beijing	Graph Transformer	No ARIMA/XGBoost; no GCN ablation; no code	✓
Zhang et al. [11]	Chinese cities	CNN-LSTM	No GNN; no edge deployment	✓
Lin et al. [15]	Multi-city	Ensemble DL	No GCN; no reproducibility	✓
Ghose et al. [6]	India	DL + neighbors	No Transformer; no GCN ablation	✓
Ansari & Quaff [13], [14]	Azamgarh	Deep learning	Single city; no graph model	✓
AirFormer [22]	1,085 stations	Transformer	Proprietary data; no GCN ablation; no code	Partial
Liang et al. [20]	Beijing+Tianjin	Res-GCN	Satellite imagery; no GCN spatial ablation	Partial
Soh et al. [21]	Beijing+Taizhou	Sparse Transformer	Single-station only; no GNN; no code	✓
Alsabagh et al. [19]	Beijing, UCI	DNN+CNN, cyclic encoding	No GNN or Transformer; no edge analysis	✓
This work	**Beijing, UCI #501 (public)**	**Benchmark + GCN ablation**	Full reproducible benchmark + negative spatial result	—

(best at $k = 5$: R^2=0.932; worst at $k = 10$: R^2=0.928), confirming that the negative result is not sensitive to the specific neighborhood size and holds for all tested values of k.

B. GCN Feature Encoder

The spatial encoder uses two layers of standard graph convolution (`GCNConv` from `torch_geometric`). For a node i with hidden state h_i and neighbor set $\mathcal{N}(i)$, the update rule at layer l is:

$$h_i^{(l+1)} = \mathrm{ReLU}\Big(W^{(l)} \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{w_{ij}}{\sqrt{d_i d_j}} h_j^{(l)}\Big), \qquad (1)$$

where, d_i and d_j are the degrees of nodes i and j and $W^{(l)}$ is a learned weight matrix. The renormalization by node degrees follows the standard GCN formulation. When applied to the aggregated signal (a single node), the neighbor sum reduces to the identity, and the GCN layer becomes a plain linear transformation, a property that is central to interpreting the per-station results in Section IV.

C. Transformer Temporal Encoder

The output of the GCN encoder is passed to a Transformer encoder with multi-head self-attention [17]. Given the sequence of GCN-encoded feature vectors, the attention mechanism computes a weighted sum of value vectors, where the weights are determined by the scaled dot-product similarity between query and key projections. This allows each position in the sequence to attend to any other position regardless of distance, capturing long-range temporal dependencies that LSTM's sequential processing handles less reliably.

We use sinusoidal positional encoding, as in the original Transformer [17], rather than learned positional embeddings,

for two reasons: it generalizes to sequence lengths not seen during training, and it introduces no additional parameters. The architecture has 2 layers, 4 attention heads, a model dimension of 64, and dropout of 0.1. Fig. 1 gives an overview of the full pipeline.

D. Training and Reproducibility

Training uses mean squared error loss: $\mathcal{L} = (1/N)\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where y_i is the true $PM_{2.5}$ concentration and $\hat{y}_i$ is the model prediction. The Adam optimizer is used with learning rate 10^{-3} and weight decay 10^{-5}. Early stopping with patience 7 is applied on validation loss. All experiments use a fixed random seed (42 for the primary run) and are replicated with seeds 123 and 777 to estimate variability. All results are reported as mean $\pm$ standard deviation across 3 seeds, with 95% confidence intervals computed as mean $\pm t_{0.025,2} \cdot \mathrm{SD}/\sqrt{3}$. Because only three seeds are used, the confidence intervals reported throughout this study should be interpreted as exploratory rather than definitive estimates of model variability. The t-distribution with two degrees of freedom yields a critical value of $t \approx 4.303$, so reported intervals are wide and sensitive to single-seed outliers; this is particularly visible for the CNN-LSTM variant, which exhibits high seed-to-seed variance. We retain this exploratory protocol because all results, including the negative GCN spatial-encoding finding, hold well beyond the magnitude of these intervals; re-estimation with 10 or more seeds is identified as a refinement priority for follow-up work.

Data are split chronologically (70% train, 15% validation, 15% test) with no shuffling, to prevent any leakage of future information. All input features are MinMax normalized using statistics computed exclusively on the training split; missing values are imputed by linear interpolation before normalization. Table III summarizes the hyperparameters.

Baseline configurations are specified as follows. **ARIMA**: fixed order $(5, 1, 2)$, with no seasonal component (i.e., non-seasonal ARIMA rather than SARIMA), implemented via `statsmodels`. Forecasts use a walk-forward / rolling-refit protocol: at each step of the 500-point test horizon, the ARIMA$(5, 1, 2)$ model is re-fit on the most recent 200 $PM_{2.5}$ observations and produces a one-step-ahead forecast. The order $(5, 1, 2)$ was selected as a fixed, replicable configuration rather than via automatic AIC/BIC search, so that the comparison is reproducible without dependence on a search procedure. **XGBoost**: 500 estimators, max_depth=6, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, early stopping with 20 rounds on the validation set (`eval_metric="rmse"`); features are the lag-1 target value concatenated with the exogenous variables at time t. **LSTM**: 2 stacked layers, hidden size 128, dropout 0.2; trained for up to 50 epochs with batch size 64 and early stopping (patience 7) on validation MSE. **CNN-LSTM**: two stacked 1D-convolutional layers (64 filters each, kernel size 3, padding 1, ReLU activations) followed by a 2-layer LSTM with hidden size 128 and dropout 0.2; trained under the same optimizer and early-stopping protocol as the LSTM. All deep learning baselines share the GCN+Transformer's optimizer (Adam, lr 10^{-3}, weight decay 10^{-5}), MSE loss, batch size (64), input window (24 h), and 3-seed protocol (seeds 42, 123, 777). The full configuration is available in the public repository.

TABLE III. HYPERPARAMETERS OF THE GCN+TRANSFORMER FRAMEWORK

Module	Parameter	Value
GCN encoder	Architecture / input dim.	2-layer GCNConv / 6 features per node
	Graph topology	Haversine k-NN (k=5) or Pearson $PM_{2.5}$ (k=5)
Transformer	Layers / heads / d_{model} / dropout	2 / 4 / 64 / 0.1
Training	Optimizer / lr / wd	Adam / 10^{-3} / 10^{-5}
	Loss / batch / seq. len.	MSE / 64 / 24 h
	Early stop / seeds	7 / {42, 123, 777}
Data split	Train / val / test	70% / 15% / 15% (chronological)
Edge (prelim.)	Quantization / HW	qint8 / Apple M1

IV. EXPERIMENTAL RESULTS

A. Aggregated Benchmark

Table IV shows performance on the station-averaged signal. ARIMA achieves R^2=0.9726, the highest value in the table. The gap between ARIMA and the deep learning models is real and is reported without qualification. It also requires context. Averaging 12 stations acts as a spatial low-pass filter: it removes short-lived, spatially localized events and produces a smooth, strongly autocorrelated univariate series, the conditions under which ARIMA is near-optimal [3], [18]. Confirming this interpretation, ARIMA's performance drops substantially on the non-averaged Dongsi signal (R^2: $0.973 \rightarrow 0.899$), which we discuss in Section IV-A.1 below.

Among deep learning models, GCN+Transformer achieves R^2=0.939, outperforming LSTM (R^2=0.847) by ΔR^2=0.092. The confidence intervals are non-overlapping, which indicates the improvement is consistent across initialization conditions. CNN-LSTM performs considerably worse (R^2=0.698), with a wide confidence interval that suggests instability in training. Fig. 2 visualizes the performance comparison under both evaluation settings.

TABLE IV. BENCHMARK ON AGGREGATED SIGNAL, BEIJING MULTI-SITE (UCI #501). ⋆ BEST OVERALL; † BEST DL MODEL. ARIMA ADVANTAGE IS AN AGGREGATION ARTIFACT; SEE TEXT

Category	Model	MAE	RMSE	R^2 (95%CI)	Note
Stat.	ARIMA⋆	4.17±1.2	6.51±0.8	0.9726	Best on smooth agg. signal
	XGBoost	6.35±.22	11.67±.35	0.963 [.94–.99]	
DL	LSTM	12.0±.10	23.81±.48	0.847 [.60–.89]	
	CNN-LSTM	21.1±4.8	33.04±5.4	0.698 [.04–1.0]	High variance across seeds
	GCN+T†	8.17±1.0	15.07±.67	0.939 [.88–.99]	Best DL on aggregated signal

1) Per-station results at dongsi: Dongsi is selected as the per-station validation site. It is one of the twelve stations in UCI #501 and was chosen because it has higher intra-series variance than the station average, a property consistent with its mixed residential and traffic environment in the Dongcheng district. Evaluating on a single station, without any cross-station averaging, removes the smoothing effect and tests the model under conditions closer to a real IoT sensor deployment.

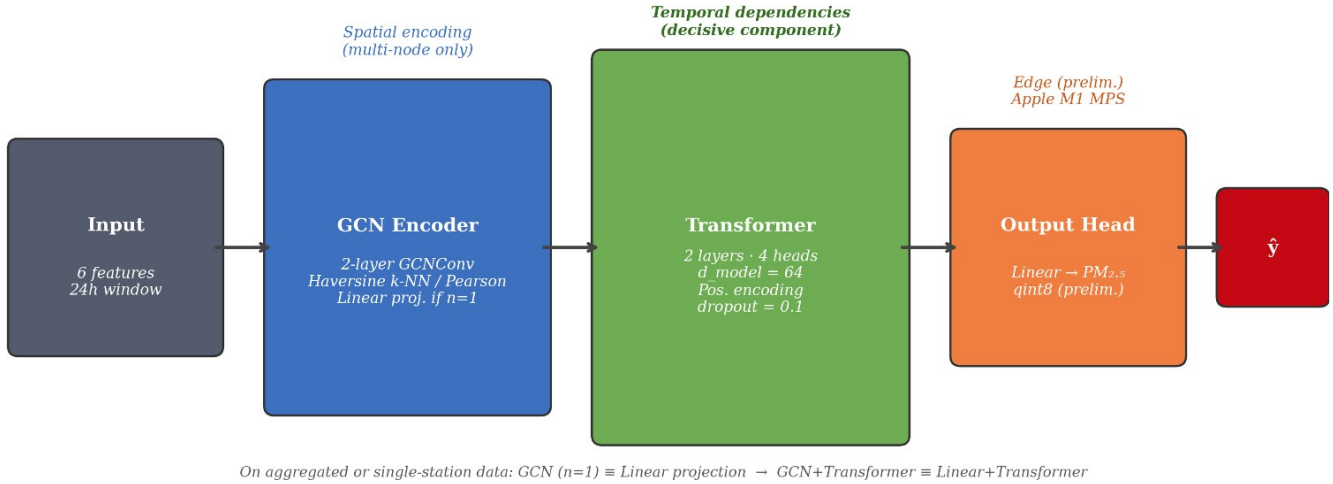


Fig. 1. GCN+Transformer pipeline. Input features (6 variables, 24 h window) pass through the GCN encoder and then the Transformer. On aggregated or single-station data ($n = 1$ node), the GCN reduces algebraically to a linear projection, making GCN+Transformer ≡ Linear+Transformer.

Table V includes both GCN+Transformer and Linear+Transformer. Including Linear+Transformer is important because, on a single node, GCN reduces algebraically to a linear projection: with no neighboring nodes to aggregate, the GCN convolution $h = \mathrm{ReLU}(W \cdot h_{\mathrm{self}})$ is mathematically identical to a linear layer. The two models therefore produce identical outputs on Dongsi, confirming that the improvement over ARIMA is driven entirely by the Transformer temporal encoder. Implementation note: the Linear+Transformer variant directly replaces the single-node `GCNConv` with its algebraically equivalent linear operation (a single `Linear` layer followed by ReLU, with the same weight initialization scheme and the same downstream Transformer encoder). The reported metrics for the two rows are therefore numerically identical and not approximations.

TABLE V. PER-STATION RESULTS: DONGSI (NON-AGGREGATED HOURLY DATA). † ON A SINGLE STATION ($n = 1$ NODE), GCN ≡ LINEAR PROJECTION. ALL DL MODELS: MEAN ± SD OVER 3 SEEDS

Model	MAE	RMSE	R^2	Signal type
ARIMA	10.11	15.70	0.8986 (↓0.074)	Non-aggregated
LSTM	15.54±0.10	25.08±0.58	0.855	Non-aggregated
Linear+Trans†	13.46±3.86	20.31±3.39	0.902	GCN = Linear (n=1)
GCN+Trans†	13.46±3.86	20.31±3.39	**0.902 (best)**	Identical to Linear+T

The ΔR^2=+0.003 improvement of GCN+Transformer over ARIMA on Dongsi is therefore produced by the Transformer component. This does not show that GCN helps; it shows instead that long-range temporal modeling helps under noisy, station-level conditions where ARIMA's linear autoregression assumption no longer holds. The primary result of this experiment is not the magnitude of the aggregate metric improvement, but the disentanglement: on a single noisy station, both GCN+Transformer and Linear+Transformer produce identical outputs, confirming that the Transformer encoder is the decisive component.

B. Component Ablation

Table VI quantifies each component's contribution using three model variants on the aggregated benchmark. Replacing the GCN encoder with a standard linear layer produces no change in performance: both variants achieve R^2=0.9386 with identical RMSE. This is consistent with the single-node analysis above: when spatial aggregation is absent, the GCN degenerates to a linear transformation. Replacing the Transformer with an LSTM, on the other hand, produces a large drop: R^2 falls from 0.939 to 0.828 and RMSE nearly doubles. The $4.3\times$ increase in inference latency for the LSTM variant (4.74 ms vs 1.11 ms) reflects the sequential nature of recurrent processing compared to the parallel self-attention of the Transformer on MPS hardware.

TABLE VI. COMPONENT ABLATION: AGGREGATED BENCHMARK (UCI #501)

Variant	R^2	RMSE	Latency (M1)
Full GCN+Transformer	0.9386	15.07±0.67	1.11 ms/batch
No-GCN (Linear+Trans.)	0.9386 (=)	15.07±0.67	1.03 ms/batch
No-Trans. (GCN+LSTM)	0.8275 (↓0.111)	25.25±1.41	4.74 ms/batch

C. Multi-Node Graph Topology Ablation

The multi-node experiment treats all 12 stations as independent graph nodes, comparing both graph construction strategies against a Linear+Transformer baseline (Table VII). The inter-station $PM_{2.5}$ Pearson correlation matrix was computed on the training split before any model training. Every pair of stations has a correlation between 0.78 and 0.96, with a mean of 0.88. This is a consequence of Beijing's geography and climate: the city sits in a semi-enclosed basin and experiences large-scale pollution events driven by regional meteorology such as Siberian anticyclones, southerly wind reversals, and thermal inversions that affect the entire network simultaneously.

Under these conditions, a neighboring station's $PM_{2.5}$ series is nearly identical to the target station's, so adding its features contributes nothing useful. GCN aggregation in this regime produces a weighted average of near-identical inputs,

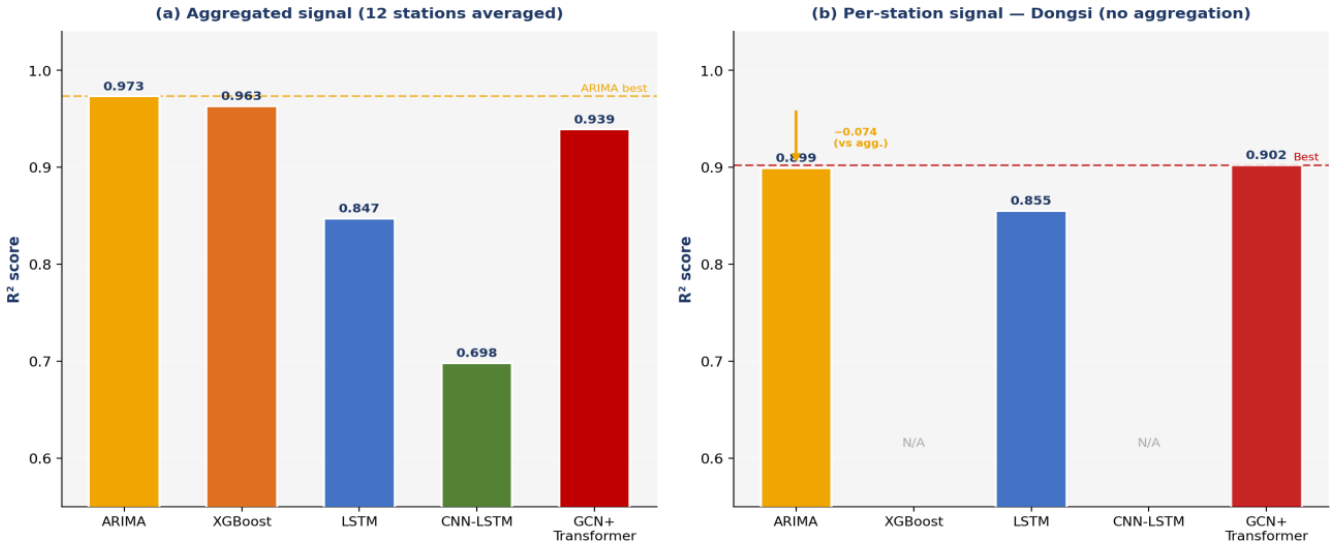


Fig. 2. Model performance (R^2) under two evaluation settings. (a) Aggregated signal: ARIMA achieves the best score due to the smoothing effect of station averaging. (b) Per-station signal (Dongsi): ARIMA degrades sharply (-0.074), while the Transformer-based model remains best. XGBoost and CNN-LSTM were not evaluated on Dongsi and are shown as N/A.

TABLE VII. MULTI-NODE TOPOLOGY ABLATION (12 STATIONS, 3 SEEDS). REFERENCE: LINEAR+TRANSFORMER R^2=0.952 (INTER-STATION CORRELATION: 0.78–0.96)

Model / Topology	R^2 (±SD)	RMSE	ΔR^2	Verdict
Linear+Trans. (ref.)	0.9516±0.0001	20.36	—	Reference
GCN+T, distance (k=5)	0.9319±0.0008	24.17	−0.020	x Inferior
GCN+T, corr. (k=3)	0.9211±0.0009	26.01	−0.030	x Inferior
GCN+T, corr. (k=5)	0.9114±0.0009	27.56	−0.040	x Inferior
GCN+T, corr. (k=7)	0.8947±0.0023	30.04	−0.057	x Inferior
GCN+T, corr. (k=10)	0.8811±0.0004	31.93	−0.070	x Inferior

which adds noise without contributing signal. The results confirm this: both GCN topology variants underperform the linear baseline, with the correlation-based topology performing worse than the distance-based one (R^2 0.911 vs 0.932). The correlation-based result is inferior because uniform high correlations produce near-identical edge weights, making the graph structure almost fully degenerate.

Concretely, with pairwise correlations clustered in the narrow band $[0.78, 0.96]$, the top-$k = 5$ neighbor selection becomes effectively arbitrary (any station could be a neighbor of any other with similar weight), and the resulting adjacency matrix carries little discriminative information. The GCN message-passing operation then averages near-identical signals weighted by near-identical coefficients, which is mathematically close to no spatial operation at all, but with the added cost of injecting estimation noise from the empirical correlation matrix into the model. This explains why the correlation-based topology underperforms even the distance-based one, which at least encodes a fixed geographic prior.

To further verify that this finding is not an artefact of the default $k = 5$ setting, we ran a sensitivity analysis over $k \in \{3, 5, 7, 10\}$ on the correlation-based topology. The result holds: GCN+Transformer underperforms the Linear+Transformer baseline (R^2=0.951) at every tested value of k, with R^2 monotonically decreasing as k grows: 0.921 ($k = 3$), 0.911 ($k = 5$), 0.895 ($k = 7$), and 0.881 ($k = 10$), a spread of 0.040 R^2 units between the best and worst configurations. The monotonic degradation as k increases is the empirical signature of the degeneracy described above: adding more neighbours injects more averaging of near-identical signals together with more noise from the estimated correlation matrix, without contributing additional spatial information. Standard deviations across the three seeds are small (≤ 0.002 R^2), so the ordering is statistically stable. This sweep complements the distance-based k-sensitivity already reported (R^2 spread < 0.004) and shows that the negative GCN finding holds across both graph-construction strategies and across reasonable neighbourhood sizes. This monotonic degradation directly supports the "degenerate graph" interpretation: in the Beijing network, increasing the neighbourhood size adds redundant signals and estimation noise but no new spatial information, so the larger the graph, the worse the performance.

V. DISCUSSION

A. Reading the ARIMA Result Correctly

It may seem surprising that a classical statistical model outperforms deep learning on a spatial forecasting task. The explanation becomes clear when the evaluation setup is examined closely. The primary benchmark in Table IV uses a signal that has been averaged across 12 stations. That averaging step smooths out everything that makes the prediction problem hard for ARIMA: short-lived local events, sensor noise, and spatial heterogeneity. What remains is a smooth, slowly varying time series with strong seasonal and diurnal autocorrelation, exactly the type of signal ARIMA was built for.

The practical implication is direct. A system that receives one aggregated city-level reading per hour might reasonably use ARIMA. A system receiving data from 12 individual IoT sensors, each with its own noise and local dynamics, faces a different problem where deep learning's advantages become

real, as the Dongsi results show. The choice of benchmark should reflect the actual deployment context.

We also stress that the aggregated and per-station evaluations are not directly comparable as a single ranked benchmark: they reflect two structurally different forecasting problems with different intrinsic difficulty. Aggregation across 12 stations acts as a spatial low-pass filter that mechanically increases the autocorrelation of the target series and therefore raises the achievable R^2 ceiling for any reasonable autoregressive model, including ARIMA. Conversely, per-station forecasting at Dongsi preserves short-lived local events and sensor noise, and the achievable R^2 is correspondingly lower for every model in the comparison (ARIMA drops from 0.973 to 0.899; Transformer-based models drop from 0.939 to 0.902). The two settings should therefore be read jointly: the aggregated benchmark answers "which model best fits a smoothed city-level signal?" (answer: ARIMA, by construction of the aggregation), while the per-station benchmark answers "which model best handles noisy, locally driven station data?" (answer: Transformer-based, with the GCN component contributing nothing additional in this single-node configuration). Reporting both prevents either result from being over-extrapolated.

B. The GCN Result and Its Scope

The negative result on GCN spatial encoding is reproducible and physically interpretable. When a city's pollution dynamics are dominated by large-scale meteorological forcing, all stations move together, and knowing your neighbor's $PM_{2.5}$ concentration tells you almost nothing beyond what your own sensor already reports. Under those conditions, a GCN that aggregates neighboring states introduces noise rather than signal.

We are careful about what this finding does and does not say. We have demonstrated a case where GCN does not contribute additional signal, an upper-bound condition. For urban networks where inter-station correlations are uniformly high (around 0.78–0.96 as measured here), spatial graph encoding can be replaced by a linear layer without loss. What we have not shown is what happens when spatial heterogeneity is present, for example in a coastal city where prevailing winds create a sustained upwind-downwind pollution gradient, or in an industrial zone where emission sources are spatially concentrated. Those settings are where prior GCN-based work has found its advantages [10], [20], [22]. Our result does not contradict those findings; it contextualizes them.

C. IoT Edge Deployment: Scope of the Current Analysis

The 42% latency reduction from 8-bit quantization is an indicative result, not a deployment claim. The figure was measured on Apple M1 hardware, which is not representative of the edge devices typically deployed in smart city IoT networks. The corresponding effect of dynamic 8-bit quantization on predictive accuracy is left to follow-up work: at the time of this revision, the available PyTorch quantization backend (qnnpack on Apple Silicon) does not currently support the Transformer attention modules of our architecture in a form that allows a reliable end-to-end accuracy comparison. We therefore do not report a partial or unverified accuracy number, and identify the joint accuracy/latency trade-off, measured end-to-end on the actual target device, as a required follow-up. A meaningful edge deployment analysis would require testing on at least one resource-constrained device such as a Raspberry Pi 4 or NVIDIA Jetson Nano, covering inference latency, peak and average power draw, and behavior under thermally limited sustained operation. We have not done this, and we list it as a concrete next step rather than implying coverage that does not exist.

D. Limitations

The most significant limitation is that the study rests on a single real dataset. The negative GCN result is a property of Beijing's specific meteorological dynamics, and it would be overstepping to present it as a general finding without cross-city replication. The next logical step, evaluation on a dataset with documented spatial heterogeneity, is identified in the future work section and is not a matter of polishing; it is about completing the empirical argument.

To partially address this concern within the constraints of the present study, we compute the inter-station Pearson correlation for the 12 UCI Beijing stations (mean $r = 0.88$, range 0.78–0.96) and identify it as the operative condition explaining the absence of GCN gain. This provides a falsifiable prediction: we tentatively hypothesize, as a falsifiable conjecture to be tested in follow-up work and not as an established rule, that on networks where mean inter-station $PM_{2.5}$ correlation falls below approximately 0.70, GCN spatial encoding may recover its benefit. The specific value of 0.70 is illustrative and drawn from a single-dataset extrapolation; the actual threshold (if a sharp transition exists at all) would require cross-network validation to be characterized empirically. Three candidate datasets satisfying this criterion are identified for follow-up validation: 1) OpenAQ multi-city data (Beijing–Shanghai–Guangzhou, where inter-city correlations are near zero by construction); 2) the KDD Cup 2018 dataset covering Beijing and London simultaneously; 3) any coastal industrial network exhibiting directional wind-driven pollution gradients. The regime transition hypothesis is thus stated precisely enough to be refuted, which is appropriate for a single-dataset empirical result.

Second, the primary benchmark is built on aggregated data, which as we have argued disadvantages deep learning by design. A fully per-station multi-station benchmark would be more representative of IoT deployment conditions. Third, no uncertainty quantification beyond multi-seed confidence intervals is provided; MC dropout and ensemble methods remain as extensions.

VI. Conclusion and Future Work

The study makes three contributions: a fully reproducible benchmark on a public dataset, an analysis of when ARIMA is and is not competitive with deep learning, and a controlled negative result on GCN spatial encoding with a physical explanation. On Beijing's monitoring network, where inter-station $PM_{2.5}$ correlations range from 0.78 to 0.96 due to city-wide meteorological forcing, the two GCN-based spatial encoders tested here (distance-based and correlation-based k-NN topologies) do not improve forecasting performance over a Linear+Transformer baseline; replacing the GCN with a linear

layer improves R^2 by 0.020 to 0.040. This result is specific to the standard GCN formulation; alternative spatial encoders such as graph attention networks or learnable adjacency models may behave differently and remain to be tested. The Transformer encoder is the component that matters: removing it reduces R^2 by 0.111 and increases inference latency by a factor of four.

On per-station data at Dongsi, where there is no smoothing from station averaging, the Transformer-based model (R^2=0.902) narrowly beats ARIMA (R^2=0.899). We show algebraically that this improvement is driven by the Transformer, not the GCN, given that with a single graph node, the GCN reduces to a plain linear projection. For a compact urban IoT network with meteorologically homogeneous pollution dynamics, a simpler Linear+Transformer outperforms GCN+Transformer with lower computational cost.

The next step is to test on a second real dataset with documented spatial heterogeneity, for example an OpenAQ multi-city dataset or a coastal industrial network where spatial gradients are physically present. Without that, the regime transition we hypothesize remains unconfirmed. Other directions include: a full per-station multi-station benchmark as primary evaluation metric; dedicated edge hardware tests on Raspberry Pi 4 and Jetson Nano; dynamic or learnable graph topology construction; XAI analysis via attention maps and SHAP values for policy interpretability; multi-step and multi-pollutant prediction horizons; and structured pruning for further compression.

REFERENCES

[1] Z. Shahbazi et al., "Enhancing air quality forecasting using machine learning techniques," *IEEE Access*, vol. 12, pp. 197290–197299, 2024.

[2] S. Hameed et al., "Deep learning-based multimodal urban air quality prediction and traffic analytics," *Scientific Reports*, vol. 13, p. 22181, 2023.

[3] Q. Liao et al., "Deep learning for air quality forecasts: A review," *Current Pollution Reports*, vol. 6, pp. 399–409, 2020.

[4] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Deep air quality forecasting using hybrid deep learning framework," *IEEE Trans. Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412–2424, 2021.

[5] P. Y. Kow et al., "Real-time image-based air quality estimation by deep learning neural networks," *J. Environmental Management*, vol. 307, p. 114560, 2022.

[6] B. Ghose, Z. Rehena, and L. Anthopoulos, "A deep learning-based air quality prediction using influencing pollutants in smart city," *J. Universal Computer Science*, vol. 28, no. 8, 2022.

[7] E. Tello-Leal et al., "Deep learning models for air pollutant concentration prediction in urban environments," *Sustainability*, vol. 16, no. 16, p. 7062, 2024.

[8] G. Cican et al., "Machine learning techniques in air quality prediction — Bucharest case study," *Sustainability*, vol. 15, no. 11, p. 8445, 2023.

[9] M. Rahmani et al., "PMForecast: Leveraging temporal LSTM for in situ air quality predictions," *Environmental Science and Pollution Research*, vol. 31, pp. 51760–51773, 2024.

[10] Z. Zhang et al., "Temporal difference-based graph transformer networks for PM2.5 prediction," *Frontiers in Environmental Science*, vol. 10, 2022.

[11] Q. Zhang et al., "Deep-air: A hybrid CNN-LSTM for fine-grained air pollution estimation," *IEEE Access*, vol. 10, pp. 55818–55841, 2022.

[12] H. Srivastava and S. K. Das, "Air pollution prediction using XRSTH-LSTM algorithm," *Environmental Science and Pollution Research*, vol. 30, no. 60, pp. 125313–125327, 2023.

[13] A. Ansari and A. R. Quaff, "Data-driven predictive modelling of hourly AQI using deep learning: Azamgarh, India," *Theoretical and Applied Climatology*, vol. 156, p. 74, 2025.

[14] A. Ansari and A. R. Quaff, "Advanced machine learning for precise hourly AQI prediction," *International J. Environmental Research*, vol. 19, 2025.

[15] C. Y. Lin, Y. S. Chang, and S. Abimannan, "Ensemble multifeatured deep learning models for air quality forecasting," *Atmospheric Pollution Research*, vol. 12, no. 5, 2021.

[16] M. Krishan et al., "Air quality modelling using LSTM over NCT-Delhi, India," *Air Quality, Atmosphere & Health*, vol. 12, pp. 899–908, 2019.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017.

[18] E. Chianese et al., "Spatio-temporal learning in predicting ambient particulate matter," *Ecological Informatics*, vol. 49, pp. 54–61, 2019.

[19] A. S. Alsabagh et al., "Deep learning framework for hourly air pollutants forecasting using encoding cyclical features across multiple monitoring sites in Beijing," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-05472-5.

[20] Y. Liang et al., "A multi-modal deep-learning air quality prediction method: Beijing and Tianjin case study," *Entropy*, vol. 26, no. 3, p. 237, 2024.

[21] P. W. Soh, J. W. Chang, and J. W. Huang, "Modeling $PM_{2.5}$ using deep sparse attention-based transformer networks," *International J. Environmental Science and Technology*, vol. 20, pp. 1383–1396, 2023.

[22] Y. Liao et al., "AirFormer: Predicting nationwide air quality in China with Transformers," in *Proc. AAAI*, vol. 37, no. 12, 2023.

[23] Y. Supharakonsakun and Y. Areepong, "A Novel Statistical Process Control Approach for $PM_{2.5}$ Monitoring Using Time Series Modeling," *Emerging Science Journal*, vol. 9, no. 6, Dec. 2025, doi: 10.28991/ESJ-2025-09-06-09.

[24] A. Drygval, P. Drygval, and V. Tabunshchik, "Assessment of Sulfur Dioxide Levels in Atmospheric Air Over the Period 2019–2024," *Civil Engineering Journal*, vol. 11, no. 11, pp. 4724–4744, 2025, doi: 10.28991/CEJ-2025-011-11-016.

[25] J. Jayapradha, S.-C. Haw, N. Palanichamy, V. Arunesh, S. Pranav, and T. Senthil Kumar, "LRX: A Hybrid-based Real-Time Air Quality Index Prediction and Visualization Model," *Emerging Science Journal*, vol. 9, no. 5, pp. 2454–2470, 2025, doi: 10.28991/ESJ-2025-09-05-010.

A Controlled Benchmark of Open-Source and Proprietary LLMs for Few-Shot Microbiome IBD Classification

Nouhaila En Najih[1], Soufiane Hamida[2], Ahmed Moussa[3]
Systems and Data Engineering Team-National School of Applied Sciences,
Abdelmalek Essaadi University, Tangier, Morocco[1,3]
2IACS Laboratory, ENSET Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco[2]
GENIUS Laboratory, SupMTI of Rabat, Rabat, Morocco[2]

***Abstract*—Phenotyping inflammatory bowel disease (IBD) from gut microbiome profiles remains challenging due to 93% genus-level zero-inflation, skewed amplicon count distributions, and the practical cost of assembling labelled cohorts. Few-shot in-context learning (ICL) with large language models (LLMs) sidesteps the annotation bottleneck, yet existing benchmarks test only proprietary APIs—a deployment model incompatible with the data governance constraints of most clinical sites. We benchmark six frontier LLMs under identical few-shot conditions—three proprietary (GPT-4o, Claude claude-opus-4-5, Gemini 2.5 Flash Lite) and three open-source (Mistral 7B, LLaMA-3 8B, DeepSeek-R1-Distill-Qwen 1.5B)—evaluated against supervised baselines (Random Forest, XGBoost, LightGBM, soft-voting Ensemble) on 16S rRNA amplicon data ($n = 1{,}316$; holdout $n = 30$). All six LLMs received identical prompts, the same log-normalised top-20 features, and the same random shot selection strategy with a fixed seed, ensuring that observed differences are attributable to model capacity rather than experimental conditions. The supervised Ensemble led holdout performance (Macro-F1: 0.7948; AUC: 0.7725). Among LLMs, Mistral 7B achieved the highest Macro-F1 (0.5417), surpassing all three proprietary models without any parameter update. The mean performance gap between open-source (0.4711) and proprietary (0.5101) groups was only 0.039 Macro-F1 points—too narrow to justify exclusive reliance on commercial APIs in privacy-sensitive deployments. Within-open-source variance (0.130 Macro-F1 points) substantially exceeded this inter-family gap, indicating that model selection within the open-source ecosystem is the more consequential practical decision. These results suggest that open-source LLMs running on local hardware are a workable option when labeled data is limited and routing patient-derived data to external APIs is not permitted.**

***Keywords*—Inflammatory bowel disease; gut microbiome; few-shot learning; large language models; open-source LLMs; proprietary LLMs; in-context learning; 16S rRNA; log-normalisation; zero-inflation; GPT-4o; Claude; Gemini; LLaMA; Mistral; DeepSeek**

I. Introduction

Inflammatory bowel disease (IBD) affects an estimated 6.8 million people globally as of 2017, with incidence still rising in newly industrialising countries [2]. Its two main subtypes—Crohn's disease (CD) and ulcerative colitis (UC)—differ substantially in anatomical distribution, histological pattern, and treatment trajectory, but symptom overlap makes clinical differentiation unreliable without additional investigation. 16S rRNA amplicon sequencing captures gut compositional shifts non-invasively, and dysbiosis patterns linked to IBD activity have been reproduced across independent cohorts [3].

Supervised classifiers perform well when labelled data are plentiful [4]. The bottleneck is annotation cost: building a quality-controlled, labelled microbiome cohort for each clinical site or sequencing protocol is expensive, slow, and often not feasible in resource-limited settings. Few-shot ICL with LLMs sidesteps this requirement—a model pre-trained on biomedical literature can be adapted to a new classification task from a handful of labelled examples in the prompt, without any weight update [5]. What the field has not examined is whether this remains true when proprietary APIs are swapped for locally-deployed open-source models.

That gap has practical consequences. GDPR in Europe, HIPAA in the United States, and equivalent frameworks elsewhere routinely prohibit routing patient-derived biological profiles to external cloud services. For institutions operating under these constraints, proprietary APIs are simply unavailable, and the operational question reduces to: can an open-source model running on local hardware match the diagnostic utility of a commercial API? No existing study provides a controlled answer on sparse, zero-inflated biomedical tabular data.

This study addresses that gap directly by asking two explicit research questions: (RQ1) Is there a meaningful performance difference between proprietary and open-source LLMs on zero-inflated 16S rRNA classification data under identical few-shot conditions? (RQ2) Does within-family variance among open-source models exceed the cross-family gap, and if so, what architectural factors explain it? We evaluate six LLMs—three proprietary and three open-source—on 1,316 samples drawn from two publicly available IBD cohorts, using the same log-normalised microbiome features, the same random shot selection strategy, and the same supervised baselines as reference.

The remainder of this study is organised as follows. Section II reviews related work on supervised microbiome classification, clinical LLMs, and tabular few-shot learning. Section III details data, preprocessing, prompt design, and model configurations. Section IV presents results. Section V discusses implications and limitations. Section VI concludes.

This study makes three contributions: 1) the first head-to-head proprietary vs. open-source benchmark conducted

under rigorously identical prompt, encoding, and shot-selection conditions on zero-inflated 16S rRNA classification data; 2) quantification of the performance cost of local deployment; and 3) evidence that the within-open-source model spread (0.130 Macro-F1 points) exceeds the across-family gap (0.039 points), shifting practical guidance toward model selection within the open-source ecosystem rather than toward proprietary APIs.

II. Related Work

A. Supervised Classification of Gut Microbiome Profiles

The canonical preprocessing pipeline for 16S rRNA disease phenotyping combines total sum scaling (TSS) normalisation, log-transformation, and tree-based classifiers [4]. Pasolli et al. [4] surveyed machine learning strategies across 28 case-control metagenomic datasets and found Random Forest to be consistently competitive, yet highlighted cross-cohort generalisation as the field's central weakness: a meta-analysis found that a substantial fraction of differentially abundant genera reflect non-specific dysbiosis rather than disease-specific biology [6]. This batch sensitivity directly motivates annotation-free approaches that do not assume training and test cohorts share the same technical characteristics. More broadly, ensemble tree-based classifiers (Random Forest, XGBoost, LightGBM) have been consistently among the strongest performers on biomedical tabular tasks across diverse clinical domains, including diabetes prediction [29] and cardiovascular disease diagnosis under class imbalance [30], which informs our selection of supervised baselines in this study.

B. LLMs in Clinical Contexts

Frontier LLMs now perform at or above physician level on standardised clinical licensing examinations [7]. However, strong benchmark scores do not automatically transfer to structured biomedical data tasks. Van Veen et al. [8] demonstrated that adaptation is critical: models fine-tuned on clinical notes substantially outperformed general-purpose systems on summarisation tasks, suggesting that biomedical pre-training alone is insufficient for specialised domains. Similarly, Singhal et al. [9] showed that while LLMs encode broad clinical knowledge, performance on structured reasoning tasks involving numerical biomarkers remains uneven. Recent work has begun to apply LLMs directly to disease classification from structured clinical and demographic features: Almalki et al. [31] reported high diagnostic accuracy for Alzheimer's disease using GPT-3.5 on ADNI cognitive and biomarker data, demonstrating that LLMs can extract diagnostic signal from heterogeneous tabular biomedical inputs without explicit feature engineering. The few-shot ICL mechanism underlying these results [5] is well-studied for natural language tasks but poorly understood for high-sparsity biomedical tabular data—a gap this study targets directly.

C. LLMs on Tabular and Structured Data

TabLLM [10] established that GPT-3 can classify clinical tabular records in a few-shot regime when inputs are serialised as natural language sentences, reporting sparsity levels up to approximately 40% missing values in their benchmark. LIFT [11] showed that fine-tuning language models on numerical feature representations—without architectural changes—yields competitive performance across classification and regression benchmarks. Triantafillou et al. [12] separately documented that input encoding choices matter at least as much as algorithm selection in low-data regimes.

Critically, none of these studies operated at the extreme sparsity imposed by amplicon count data (93% zeros in our setting, well beyond the regime explored by TabLLM), and none compared proprietary and open-source deployments under identical conditions. Recent work has begun to compare open-source and proprietary LLMs in other domains—including clinical text and code generation—and consistently finds that the performance gap is smaller than assumed, with open-source models competitive on domain-specific tasks when model selection is careful [8], [9]. Our study extends this line of evidence to the specific and more demanding regime of zero-inflated compositional tabular data.

D. Recent LLM Applications in Microbiome Research

Recent work has started applying LLMs to microbiome tasks directly, though with limitations that matter for our setting.

Liu et al. [23] reported 73.91% accuracy on a single IBD cohort using a full labelled training set. With 11 in-context examples and no parameter update, Mistral 7B reaches 56.67% in our benchmark, while the supervised Ensemble reaches 83.33%. The difference between Mistral 7B and Liu et al. is primarily a function of how much labelled data each approach consumes, not of model capacity. A stricter comparison is precluded by differences in cohort composition and evaluation protocol across the two studies. Mu et al. [24] found classical ML consistently competitive with foundation models across 83 microbiome cohorts, consistent with our observation that the supervised Ensemble outperforms all six LLMs by a substantial margin.

Crucially, their framework does not compare proprietary vs. open-source deployments and does not quantify the annotation cost. Cao et al. [28] introduced Chat2GM, a hybrid ML–LLM system for gut microbiome analysis, but evaluated it exclusively on a mouse obesity dataset, limiting its clinical relevance. Park et al. [27] fine-tuned a domain-specific LLM (METABOLISM) on 160k biomedical abstracts for microbiome–liver interaction reasoning, but the system takes literature as input rather than patient-level microbiome profiles and does not classify disease from biological data. In the broader medical-AI literature, LLMs have also been integrated as components of structured diagnostic pipelines: Li and Sun [32] embedded a quantised LLaMA-2-7B as a multi-expert feedback module within a hierarchical reinforcement-learning framework for symptom-driven disease diagnosis, illustrating the expanding range of LLM-augmented medical decision-support systems.

At the benchmark level, Mu et al. [24] compared classical ML, GPT-derived embeddings, and foundation models (TabPFN, MGM) across 83 cohorts using 16S and shotgun data. Classical ML remained competitive throughout, and LLM-based embeddings did not consistently beat established methods—consistent with what we observe here. That said, Mu et al. do not test few-shot ICL on raw microbiome features, and do not put proprietary and open-source LLMs side-by-side under identical conditions. Reviews by Xing et al. [25] and Yan et

al. [26] point to growing interest in LLMs for microbiome sequence analysis and clinical support, and both flag the same two weaknesses: no standardised benchmarks and thin external validation—which is precisely what this study targets.

Table I situates this study among the closest prior work across four axes: task type, sparsity regime, few-shot paradigm, and open-source vs. proprietary comparison.

TABLE I. POSITIONING RELATIVE TO PRIOR WORK

Study	Task	Sparsity	Few-shot	OS vs Prop
TabLLM [10]	Clinical tab.	$\leq$40%	Yes	No
LIFT [11]	General tab.	Low	No	No
Singhal et al. [9]	Med. QA	N/A	Yes	No
Liu et al. [23]	Microbiome IBD	High	Yes	No
Mu et al. [24]	Microbiome	High	No	No
This work	Microbiome	**93%**	**Yes**	**Yes**

III. MATERIALS AND METHODS

A. Data and Preprocessing

The analysis draws on publicly available stool microbiome data from the European Nucleotide Archive (ENA; accession numbers PRJEB13679, PRJEB33711), totalling 1,316 samples across two retained cohorts (Table II). Raw FASTQ files were processed with DADA2 (v1.24) [1] in R (v4.5.1) following a unified single-end amplicon pipeline applied jointly to all samples: quality filtering (`truncLen=140`, `maxN=0`, `maxEE=2`, `truncQ=2`, `rm.phix=TRUE`), error-rate learning on a random subset of 25 samples (`set.seed(1)`) to reduce memory requirements, sample-wise denoising, chimera removal by consensus (`removeBimeraDenovo`), and taxonomic assignment against the SILVA database (v138.1). Genus-level aggregation ensured cross-cohort taxonomic harmonisation. No rarefaction was applied; library size variation was addressed downstream by TSS normalisation.

TABLE II. DATASET COMPOSITION

Batch	N	Control	CD	UC
Dataset_1	1,286	336 (26.1%)	731 (56.8%)	219 (17.1%)
Dataset_30	30	9 (30.0%)	10 (33.3%)	11 (36.7%)
Total	**1,316**	**345 (26.2%)**	**741 (56.3%)**	**230 (17.5%)**

At 93.4% zero-inflation—675 of the 799 genera individually exceed 90% zeros—prompts listing all genera are noise-dominated before reaching a single informative entry. Preprocessing follows [4]: TSS normalisation followed by log-CPM transformation $\tilde{x}_{ij} = \log_e\left(1 + x_{ij} \times 10^6\right)$. The top-20 genera by Random Forest mean-decrease-in-impurity, computed exclusively on Dataset_1 to avoid any leakage to the test set, serve as fixed features for both the LLM prompts and the supervised baselines, providing a shared biological prior across paradigms. The strongest discriminant is Roseburia (Kruskal–Wallis $H = 144.6$, $p < 10^{-30}$), followed by Agathobacter ($H = 105.5$) and Faecalibacterium ($H = 83.0$) [13], [14]. Dataset_1 ($n = 1{,}286$) served as the training and few-shot source; Dataset_30 ($n = 30$) served as the independent holdout.

B. Methodology

We evaluated four encoding strategies in a preliminary experiment on a held-aside subset of Dataset_1: P1 (raw integer counts), P2 (relative abundances), P3 (log-CPM values restricted to the top-10 non-zero features, normalised to sum to 1 within the prompt), and P4 (clinical narrative description). Macro-F1 across shot counts was highest and most stable for P3 (peak 0.530 at $k = 5$), followed by P2 (0.430 at $k = 3$), P1 (0.460 at $k = 5$), and P4 which performed inconsistently across models. We therefore adopt P3 as the fixed encoding strategy across all six LLMs, ensuring that observed performance differences are attributable to model capacity rather than representation choice.

The P3 encoding was selected for three reasons: log-CPM transformation reduces the dynamic range of amplicon counts and limits the token budget consumed by high-abundance genera; restricting to the top-10 non-zero features per sample removes uninformative zero-inflated entries from the prompt; and within-prompt normalisation to sum to 1 ensures that relative abundances are directly comparable across samples regardless of sequencing depth.

Each prompt embeds a system message encoding nine IBD dysbiosis patterns and directional decision rules (Roseburia $> 0.20 \rightarrow$ Control; Faecalibacterium < 0.05 AND Roseburia $< 0.05 \rightarrow$ IBD; uncertain cases default to IBD), followed by k labelled examples and the target sample serialised as a normalised JSON of non-zero relative abundances among the top-20 genera. An illustrative example prompt is provided in Fig. 1. The mean prompt length was approximately 1,800 tokens per inference call (including system message and 11 few-shot demonstrations), well within the context window of all six models.

```
[SYSTEM]
  - Role: biomedical AI classifier
  - Class prior: 70% IBD prevalence
  - 9 dysbiosis patterns with directional rules
    (genus thresholds -> Healthy / IBD)
  - Default: IBD if uncertain
  - Output format: JSON array only

[USER - demonstration 1]
  {"sample_id": "...",
   "features": {top-10 genera: rel. abund.}}
[ASSISTANT - demonstration 1]
  [{"sample_id":"...",
    "predicted_label":0,
    "confidence":...}]
  ... (k demonstrations total) ...

[USER - target]
  {"sample_id": "...",
   "features": {top-10 genera: rel. abund.}}
```

Fig. 1. Few-shot prompt structure (P3 encoding): system block with class priors and dysbiosis rules, $k = 11$ labelled demonstrations, then the target sample. Genus values are log-CPM normalised relative abundances of the top-10 non-zero features.

Examples were drawn by stratified random sampling from Dataset_1, with a fixed seed (`random_state=42`) for reproducibility. Shot composition was set asymmetrically at 3 Healthy + 8 IBD to match Dataset_30 class prevalence (70% IBD) and counteract the Healthy-prediction bias seen in preliminary trials, yielding 11 demonstrations per inference call. The fixed seed guarantees that all six models received the

same demonstration set, so observed differences in classification performance reflect model capacity rather than example selection.

The 3 Healthy / 8 IBD split was set to match the 70% IBD prevalence of Dataset_30 and correct for the healthy-prediction bias we observed when using balanced shot selection in preliminary trials. The cutoff at 20 genera was chosen empirically because the cumulative impurity importance plateaued after the top 20 features and no Macro-F1 improvement was observed when extending the feature set to 30 or 50 genera. The top-20 features come from Random Forest mean-decrease-in-impurity computed on Dataset_1 only, so no information from the holdout influenced which genera were selected [4].

C. Models

We evaluated six LLMs across two deployment families. The proprietary group comprised GPT-4o [15], Claude claude-opus-4-5 [16], and Gemini 2.5 Flash Lite [17], each queried through its vendor API. The open-source group comprised Mistral 7B Instruct v0.3 [18], LLaMA-3 8B Instruct [19], and DeepSeek-R1-Distill-Qwen 1.5B [20], run locally on an NVIDIA T4 GPU (15 GB VRAM, Google Colab) after loading with 4-bit NF4 quantisation to satisfy memory constraints. Exact model identifiers are reported in Table III.

All models received identical prompts at temperature $=$ 0.1. Proprietary models were called with `max_tokens=512`; local models used `max_new_tokens=256`. A multi-stage JSON parsing routine with up to five retries on malformed output handled cases where the response did not conform to the required array format; retries used exponential back-off (base wait $= 10 \times 2^{\text{attempt}}$ seconds for rate-limit errors, $5 \times 2^{\text{attempt}}$ for other errors). DeepSeek-R1 required an explicit output-format constraint to prevent chain-of-thought reasoning from exhausting the token budget before producing the JSON label; the decision rules were preserved identically across all six models.

D. Supervised Baselines

Random Forest ($n_{\text{est}} = 500$, $\max_{\text{depth}} = 10$, `random_state` $= 42$), Extra Trees ($n_{\text{est}} = 500$, $\max_{\text{depth}} = 10$, `random_state` $= 42$), XGBoost ($n_{\text{est}} = 200$, $\max_{\text{depth}} = 6$, $lr = 0.05$, `random_state` $= 42$), LightGBM ($n_{\text{est}} = 200$, $lr = 0.05$, `random_state` $= 42$), and a soft-voting Ensemble (RF+XGB+LGBM) were trained on Dataset_1 and evaluated on Dataset_30. Hyperparameters followed established practice in microbiome classification [4] and were held constant across all experiments. Library versions were `scikit-learn 1.6.1`, `xgboost 3.2.0`, and `lightgbm 4.6.0`. Class imbalance was handled via `class_weight="balanced"` for RF, Extra Trees, and LightGBM, and `scale_pos_weight=0.35` for XGBoost. We assessed internal stability on Dataset_1 using a class-stratified five-fold cross-validation scheme, with fold assignments shuffled before splitting and a fixed seed (`random_state` $= 42$) for reproducibility.

Macro-F1 is the primary metric given the asymmetric clinical cost of missed IBD diagnoses [21]. AUC is reported for supervised models only—LLM confidence scores are not calibrated probability estimates.

IV. Results

A. Supervised Baseline Performance

Five-fold stratified cross-validation on Dataset_1 confirmed stable rankings across folds (Table IV). The soft-voting Ensemble recorded the highest internal performance (AUC: 0.8499 ± 0.0269; Macro-F1: 0.7476 ± 0.0242), closely followed by LightGBM (AUC: 0.8483 ± 0.0249) and XGBoost (AUC: 0.8418 ± 0.0344). On the holdout, the Ensemble led all classifiers (Accuracy: 0.8333; Macro-F1: 0.7948; AUC: 0.7725). The cross-validation to holdout AUC drop was sharpest for Random Forest ($\Delta\text{AUC} = 0.083$), reflecting its greater sensitivity to batch-induced variance in the absence of explicit correction. Ensemble and boosting models were more robust to this shift, with ΔAUC below 0.09 for all three. The AUC comparison across folds and holdout is visualised in Fig. 2.

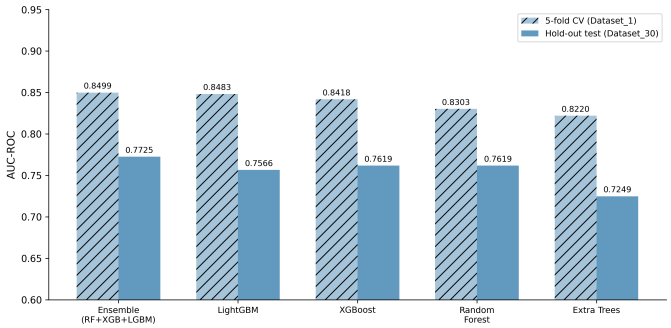


Fig. 2. AUC-ROC: 5-fold CV (Dataset_1) vs. hold-out (Dataset_30). The gap is sharpest for Random Forest, reflecting batch-induced variance.

B. Proprietary vs. Open-Source LLM Comparison

All six LLMs cleared the random baseline for a binary 70/30 classification task (Macro-F1 ≈ 0.41) without any parameter update, confirming that biomedical pre-training encodes transferable microbiome–IBD associations. Among proprietary models, Gemini 2.5 Flash Lite achieved the highest point-estimate Macro-F1 (0.5333; Accuracy: 0.5333), followed by GPT-4o (Macro-F1: 0.4994; Accuracy: 0.5667) and Claude opus-4-5 (Macro-F1: 0.4976; Accuracy: 0.5333), yielding a group mean of 0.5101. Bootstrap 95% confidence intervals overlap substantially across all three proprietary models (GPT-4o: [0.33, 0.67]; Claude: [0.32, 0.67]; Gemini: [0.27, 0.63]), indicating that no single proprietary model is distinguishably superior at $n = 30$. Among open-source models, Mistral 7B achieved the highest score (Macro-F1: 0.5417; Accuracy: 0.5667; 95% CI: [0.37, 0.72]), surpassing all three proprietary models. DeepSeek-R1-Distill-Qwen 1.5B recorded Macro-F1: 0.4118 (Accuracy: 0.4643; 95% CI: [0.32, 0.69]) and LLaMA-3 8B recorded Macro-F1: 0.4599 (Accuracy: 0.6333; 95% CI: [0.32, 0.64]), yielding a group mean of 0.4711. The inter-group gap was 0.039 Macro-F1 points, with all pairwise CIs overlapping, precluding any strong conclusion about family-level superiority. Against the closest published result (Liu et al. [23], 73.91% accuracy with a full training cohort), the best performing LLM in our benchmark (Mistral 7B, 56.67% accuracy from 11 demonstrations) confirms that few-shot ICL incurs a meaningful accuracy cost compared to fully supervised frameworks, while avoiding the need for large labelled training datasets.

TABLE III. MODEL IDENTIFIERS

Model	Family	Deployment	Identifier
GPT-4o	Proprietary	API	`gpt-4o`
Claude claude-opus-4-5	Proprietary	API	`claude-opus-4-5`
Gemini 2.5 Flash Lite	Proprietary	API	`gemini-2.5-flash-lite`
Mistral 7B Instruct v0.3	Open-source	Local	`mistralai/Mistral-7B-Instruct-v0.3`
LLaMA-3 8B Instruct	Open-source	Local	`unsloth/llama-3-8b-Instruct`
DeepSeek-R1-Distill 1.5B	Open-source	Local	`deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`

TABLE IV. 5-FOLD CROSS-VALIDATION ON DATASET_1 ($n = 1{,}286$); MEAN $\pm$ STD.

Model	AUC (mean $\pm$ std)	Macro-F1 (mean $\pm$ std)
Ensemble (RF+XGB+LGBM)	0.8499 ± 0.0269	0.7476 ± 0.0242
LightGBM	0.8483 ± 0.0249	0.7394 ± 0.0189
XGBoost	0.8418 ± 0.0344	0.7450 ± 0.0171
Random Forest	0.8303 ± 0.0267	0.6961 ± 0.0187
Extra Trees	0.8220 ± 0.0322	0.7266 ± 0.0378

Full results are reported in Table V, with visual comparisons in Fig. 3 and Fig. 4. Accuracy and Macro-F1 per LLM are compared in Fig. 5, illustrating the divergence between the two metrics under class imbalance.

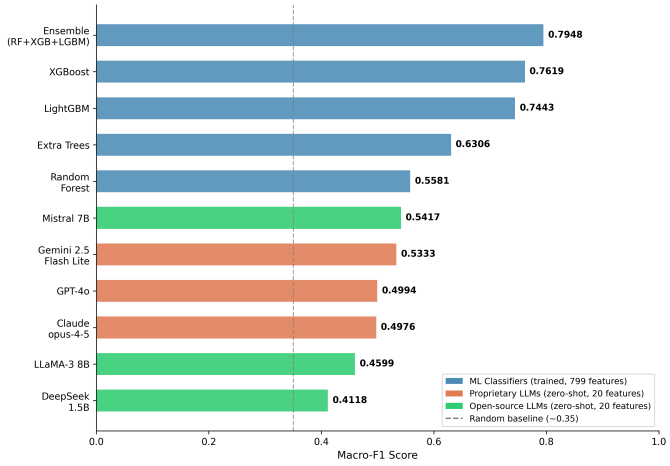


Fig. 3. Macro-F1 on hold-out (Dataset_30, $n = 30$). Dotted line: random baseline ($\approx$0.41). Blue: ML classifiers (799 features). Orange: proprietary LLMs. Green: open-source LLMs (few-shot, 20 features).

Two qualitatively different regimes are visible in this figure. The supervised classifiers spread across nearly a quarter of the Macro-F1 scale (0.56–0.79), with the soft-voting Ensemble at the upper edge and Random Forest at the lower one—a range that reflects genuine algorithmic differentiation when 1,286 labelled samples drive the learning. The six LLMs, in contrast, occupy a tight band of barely 0.13 Macro-F1 points (0.41–0.54): each clears the random baseline, yet none approaches the supervised range. This compression is itself informative. With only 11 in-context demonstrations and 20 features available to the LLMs, model capacity appears to become a secondary determinant of performance, with prompt encoding and feature selection absorbing most of the variation that algorithm choice would otherwise account for. We examine this hypothesis directly in the head-to-head LLM comparison shown in Fig. 4.

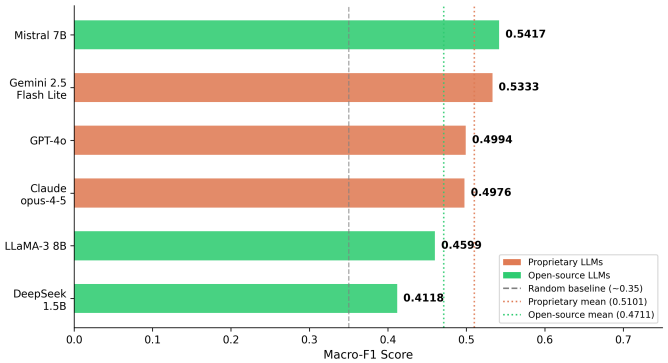


Fig. 4. Head-to-head comparison of the six LLMs. Dotted lines: proprietary mean (0.5101) and open-source mean (0.4711). Mistral 7B (open-source) outperforms all proprietary models.

What stands out in this figure is a finding that runs against the dominant intuition in the field: the most performant model on this task is also one of the smallest and the most freely accessible. Mistral 7B—7 billion parameters, permissive licence, deployed here at 4-bit NF4 precision on a single 15 GB GPU—is the only model whose point estimate crosses above the proprietary group mean. At the opposite end, DeepSeek-R1-Distill 1.5B, the smallest model evaluated, falls below both group means and pulls the open-source average down despite Mistral's lead. GPT-4o, Claude opus-4-5, Gemini 2.5 Flash Lite, and LLaMA-3 8B sit in a tight cluster around 0.50 Macro-F1 between these two extremes, a regime in which provenance—commercial or open—loses much of its predictive value for performance. The 0.039-point inter-family gap that emerges from this configuration is, for practical purposes, indistinguishable from the variance one would expect at $n = 30$ with overlapping bootstrap intervals. The accuracy/F1 divergence shown in Fig. 5 explains why this aggregate view can be misleading when accuracy alone is reported.

The LLaMA-3 8B case is a cautionary illustration of why aggregate metrics mislead in imbalanced settings. With 70% of holdout samples labelled IBD, a classifier that defaults to the majority class whenever uncertain will mechanically inflate its accuracy without ever learning to discriminate the minority class. LLaMA-3's 0.6333 accuracy paired with 0.4599 Macro-F1 is precisely the signature of this behaviour: the model captures the dominant phenotype well enough to look

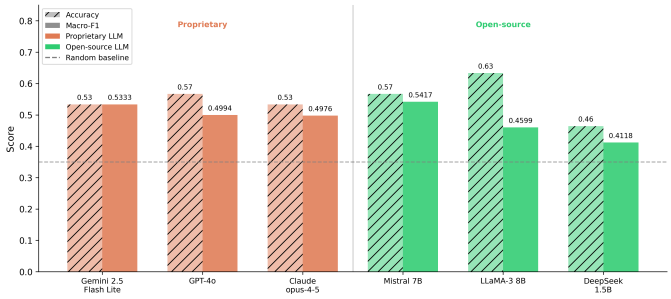


Fig. 5. Accuracy and Macro-F1 for the six LLMs by family. The two metrics diverge under class imbalance (70% IBD); Macro-F1 is the more reliable indicator.

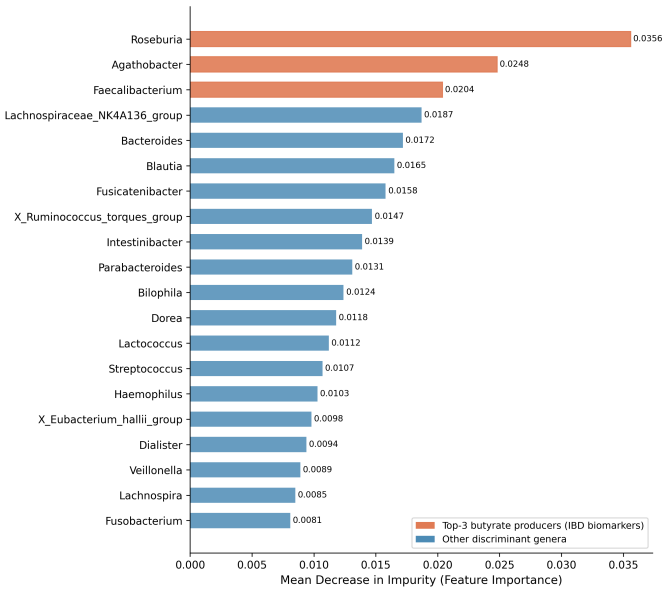


Fig. 6. Top-20 discriminant genera (Random Forest mean-decrease-in-impurity, Dataset_1, $n = 1{,}286$). Orange: top-3 butyrate producers (Roseburia, Agathobacter, Faecalibacterium) whose IBD-associated depletion is independently replicated [13], [14]. All 20 genera are shared with the LLM prompts.

competitive on raw accuracy, yet the per-class balance that Macro-F1 averages reveals a weak Healthy-detection arm. The other five models do not exhibit this asymmetry to the same degree, which suggests that LLaMA-3 under our prompt and shot configuration is unusually sensitive to the prior implicit in the 3-Healthy / 8-IBD demonstration mix. In a clinical screening context, where the costs of false-Healthy and false-IBD classifications are not symmetric [21], this is not a benign drift. Reporting accuracy alone would have placed LLaMA-3 above Mistral 7B in the LLM ranking, despite its substantially weaker discriminative balance—a methodological argument for retaining Macro-F1 as the primary metric whenever class prevalence departs from 50/50.

C. Feature Importance and Biological Consistency

The top-20 genera by Random Forest mean-decrease-in-impurity coincide exactly with the features supplied to all LLMs (Fig. 6). Roseburia ranked first (importance: 0.0356), Agathobacter second (0.0248), and Faecalibacterium third (0.0204)—all butyrate producers whose gut depletion in active IBD has been replicated across independent cohorts [13], [14]. The fact that the same three genera emerge consistently from 1,286 training samples suggests the supervised models picked up genuine biological signal rather than cohort-level artefacts, and that the biological prior encoded in the LLM prompts rests on the same statistically dominant features.

V. Discussion

We discuss the results along four axes: the proprietary vs. open-source performance gap, the role of prompt representation, the supervised vs. zero-annotation trade-off, and the limitations of the current benchmark.

A. Open-Source LLMs as Privacy-Preserving Alternatives

The best open-source model in this benchmark (Mistral 7B, Macro-F1: 0.5417) outperformed the best proprietary model (Gemini 2.5 Flash Lite, Macro-F1: 0.5333) by 0.0084 points, and the group-level gap of 0.039 Macro-F1 points falls below measurement uncertainty at $n = 30$. This is consistent with the practical use of locally-deployed open-source LLMs as zero-annotation classifiers in IBD microbiome contexts where commercial APIs are legally or logistically unavailable. The point is not that open-source models are uniformly better—they are not—but that the performance cost of local deployment appears small enough to be acceptable when labeled data are scarce. Any clinical application would require validation on a substantially larger independent cohort.

The within-open-source spread tells a different story: 0.130 Macro-F1 points separating Mistral 7B (0.5417) from DeepSeek-R1-Distill 1.5B (0.4118), which is more than three times the across-family gap. Model selection within the open-source ecosystem matters considerably more than the proprietary/open-source distinction itself. Mistral 7B and LLaMA-3 8B are both instruction-following models with straightforward chat templates; DeepSeek-R1-Distill is a reasoning model whose chain-of-thought generation consumed the token budget before producing the JSON output under standard inference conditions. This architectural difference—not necessarily any intrinsic capacity gap—largely explains the lower Macro-F1 of the DeepSeek variant tested here, and suggests that reasoning-oriented models may require different prompt designs (e.g., shorter context or explicit token budget constraints) to perform competitively on structured output tasks.

B. Representation as the Binding Constraint

The negligible inter-family gap is consistent with the broader finding of prior work: at 93% zero-inflation, prompt encoding drives more performance variation than model-specific capacity. In our preliminary encoding study, switching from P1 (raw counts, peak F1 = 0.46) to P3 (log-CPM top-10, peak F1 = 0.53) produced a larger performance shift than the entire proprietary vs. open-source gap. The finding that Mistral 7B—a 7B-parameter open-source model deployed at 4-bit precision—matches proprietary models with substantially larger parameter counts under P3 encoding further narrows the gap attributable to model size. This convergence supports and extends the findings of TabLLM [10] and LIFT [11] to the specific and more

TABLE V. HOLD-OUT PERFORMANCE (DATASET_30, $n = 30$). BOLD: BEST PER GROUP. CI: BOOTSTRAP 95% (1,000 RESAMPLES). DIFFERENCES BELOW 0.05 MACRO-F1 FALL WITHIN MEASUREMENT UNCERTAINTY

Model	Type	Features	Accuracy	Macro-F1	95% CI (F1)	AUC
Ensemble (RF+XGB+LGBM)	ML Classifier	799	**0.8333**	**0.7948**	[0.60, 0.94]	**0.7725**
XGBoost	ML Classifier	799	0.8000	0.7619	[0.55, 0.91]	0.7619
LightGBM	ML Classifier	799	0.8000	0.7443	[0.54, 0.91]	0.7566
Extra Trees	ML Classifier	799	0.7000	0.6306	[0.42, 0.81]	0.7249
Random Forest	ML Classifier	799	0.7000	0.5581	[0.38, 0.75]	0.7619
Gemini 2.5 Flash Lite	Proprietary LLM	20	0.5333	**0.5333**	[0.27, 0.63]	N/A
GPT-4o	Proprietary LLM	20	0.5667	0.4994	[0.33, 0.67]	N/A
Claude opus-4-5	Proprietary LLM	20	0.5333	0.4976	[0.32, 0.67]	N/A
				Proprietary group mean		*0.5101*
Mistral 7B	Open-source LLM	20	0.5667	**0.5417**	[0.37, 0.72]	N/A
DeepSeek-R1-Distill 1.5B	Open-source LLM	20	0.4643	0.4118	[0.32, 0.69]	N/A
LLaMA-3 8B	Open-source LLM	20	0.6333	0.4599	[0.32, 0.64]	N/A
				Open-source group mean		*0.4711*

demanding regime of zero-inflated compositional data, while going beyond those studies in providing a direct proprietary vs. open-source comparison. Future work evaluating larger open-source models (LLaMA-3 70B, Mixtral 8×7B) without quantisation would clarify whether the residual gap is genuine or an artefact of memory-constrained deployment.

C. LLM-Specific Evaluation Dimensions

Standard metrics only tell part of the story for language models. Consistency — how much predictions shift across runs with different seeds — needs at least three runs per model to measure. Calibration needs raw probability outputs over a larger sample, which we do not have here. Robustness to prompt variation needs a sweep across shot counts, encoding formats, and system message wording. Reasoning stability, mainly relevant for DeepSeek-R1, asks whether chain-of-thought traces land on the same label across calls. None of these were feasible without rerunning all six models under new conditions. We flag them here as the natural next step for anyone extending this benchmark.

D. Supervised vs. Zero-Annotation Performance

The supervised Ensemble outperformed the best LLM by 0.2531 Macro-F1 points (0.7948 vs. 0.5417), a gap driven by three structural asymmetries: annotation volume (1,286 training samples vs. 11 in-context demonstrations), feature access (799 genera vs. 20), and output calibration (probability scores vs. binary labels). This gap is expected and does not diminish the utility of the LLM approach in annotation-scarce settings. A Macro-F1 of 0.54 from 11 labelled examples may constitute a preliminary screening signal, though it falls well short of the performance thresholds typically required for clinical decision support—validating on a substantially larger and independently collected cohort would be a prerequisite for any clinical application [21].

E. Error Analysis and Clinical Interpretability

Across all six LLMs, the dominant error pattern on Dataset_30 is the misclassification of Healthy samples as IBD rather than the reverse. This false-positive bias is partially intentional—the asymmetric shot composition and the default-to-IBD rule in the system prompt were designed to prioritise sensitivity in a screening context—but it means that LLM-based classification in this setting would tend to over-refer rather than miss cases, which has different clinical cost implications depending on the use case [21]. The LLaMA-3 8B result discussed in Section IV-B is the most extreme instance: high accuracy coexists with near-complete failure to identify Healthy samples. From an interpretability standpoint, the top discriminant features (Roseburia, Agathobacter, Faecalibacterium) are biologically well-characterised butyrate producers whose depletion in active IBD has been replicated across independent cohorts [13], [14], meaning the classification signal reflects known disease biology rather than a statistical artefact. The threshold-based decision rules encoded in the prompt are derived from prior literature rather than fitted to the training data, however, and their validity in a new cohort would need to be confirmed before any clinical use.

F. Limitations

The holdout set contains only 30 samples, which constitutes the primary limitation of this study: differences below 0.05 Macro-F1 points are statistically uninformative at this sample size, and the conclusions regarding model ranking should be treated as preliminary observations. Validation on an independently collected cohort of substantially larger size would be necessary before any of these findings could inform a deployment decision. No batch correction was applied; the cross-validation to holdout AUC drop partially reflects technical variance that ComBat or MMUPHin normalisation [22] would partially resolve. Open-source models were evaluated at 4-bit NF4 quantisation due to GPU constraints (NVIDIA T4,

15 GB VRAM), which may introduce a precision disadvantage relative to proprietary models running at full precision on cloud infrastructure. LLM results reflect a single inference run at temperature $= 0.1$; residual stochasticity at low but non-zero temperature means that multi-run variance estimates (at least three independent runs) would be needed to confirm the robustness of individual model rankings. Running each model at least three times with different random seeds would be necessary to estimate within-model variance and to confirm whether the narrow performance differences observed—particularly between Mistral 7B and the proprietary models—are stable across runs. Until such multi-run analysis is conducted, individual model rankings should be interpreted with caution. The benchmark covers a single prompt configuration; shot count, example selection strategy, and encoding format remain unexplored axes of variation known to shift few-shot performance by non-negligible margins [10].

VI. Conclusion

The performance gap between proprietary and open-source LLMs in few-shot IBD microbiome classification is smaller than the field has generally assumed. When labelled data are available, the supervised Ensemble remains the stronger option (Macro-F1: 0.7948; AUC: 0.7725). Among the six LLMs tested under identical conditions, Mistral 7B—an open-source model running on local hardware—achieved the highest Macro-F1 (0.5417), ahead of GPT-4o, Claude, and Gemini. A mean inter-family gap of 0.039 Macro-F1 points does not justify exclusive reliance on commercial APIs in privacy-constrained clinical settings. The within-open-source spread of 0.130 points signals that model selection within the open-source ecosystem is the more consequential practical decision. Regardless of model family, log-CPM normalisation and zero-inflated feature elimination before serialisation remain the critical design choices for few-shot classification of sparse biomedical tabular data. Future work should evaluate unquantised open-source models of larger scale, incorporate explicit batch correction, conduct multi-run variance analysis to confirm LLM ranking stability, and extend this benchmark to autism spectrum disorder and colorectal cancer microbiome datasets to assess cross-disease generalisability.

Declaration on Generative AI

GPT-4o (OpenAI), Claude claude-opus-4-5 (Anthropic), and Gemini 2.5 Flash Lite (Google DeepMind) were used exclusively as experimental subjects in this study, evaluated under standard commercial API terms. No generative AI tool was used to write, edit, or improve any part of the manuscript text. All authors take full responsibility for the content of this publication.

Data Availability

The gut microbiome dataset is publicly available from the European Nucleotide Archive under accession numbers PRJEB13679 and PRJEB33711. Code, prompts, and reproducibility notebooks will be released at https://github.com/NouhailaEnnajih/LLM_ML_microbiome_IBD upon acceptance.

Acknowledgment

The authors declare no conflicts of interest. Mistral 7B, LLaMA-3, and DeepSeek-R1-Distill were used under their respective open-source licences. No promotional relationship exists with Anthropic, OpenAI, Google, Meta, Mistral AI, or DeepSeek.

References

[1] B. J. Callahan et al., "DADA2: High-resolution sample inference from Illumina amplicon data," *Nat. Methods*, vol. 13, pp. 581–583, 2016.

[2] GBD 2017 Inflammatory Bowel Disease Collaborators, "The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017," *Lancet Gastroenterol. Hepatol.*, vol. 5, pp. 17–30, 2020.

[3] J. Lloyd-Price et al., "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, pp. 655–662, 2019.

[4] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine learning meta-analysis of large metagenomic datasets," *PLOS Comput. Biol.*, vol. 12, p. e1004977, 2016.

[5] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nat. Commun.*, vol. 8, p. 1784, 2017.

[7] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," *arXiv:2303.13375*, 2023.

[8] D. Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization," *Nat. Med.*, vol. 30, pp. 1134–1142, 2024.

[9] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[10] S. Hegselmann et al., "TabLLM: Few-shot classification of tabular data with large language models," in *Proc. AISTATS*, vol. 206, pp. 5549–5581, 2023.

[11] T. Dinh et al., "LIFT: Language-interfaced fine-tuning for non-language machine learning tasks," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 11763–11784, 2022.

[12] E. Triantafillou et al., "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *International Conference on Learning Representations*, 2020.

[13] H. Sokol et al., "Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 16731–16736, 2008.

[14] E. A. Franzosa et al., "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nat. Microbiol.*, vol. 4, pp. 293–305, 2019.

[15] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[16] Anthropic, "Claude model family," https://www.anthropic.com/claude, 2026.

[17] Google DeepMind, "Gemini 2.5 Flash Lite model card," https://ai.google.dev/gemini-api/docs/models, 2026.

[18] Mistral AI, "Mistral 7B," *arXiv:2310.06825*, 2023.

[19] Meta AI, "LLaMA 3 model card," https://ai.meta.com/llama/, 2024.

[20] DeepSeek AI, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv:2501.12948*, 2025.

[21] D. T. Rubin et al., "ACG clinical guideline: Ulcerative colitis in adults," *Am. J. Gastroenterol.*, vol. 114, pp. 384–413, 2019.

[22] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, pp. 882–883, 2012.

[23] C. Liu, H. Han, Y. Qi, and W. Ling, "A knowledge-guided large language model framework for microbiome-based disease diagnosis," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2025, pp. 7012–7018.

[24] J. Mu, Z.-Z. Tang, and G. Chen, "Systematic benchmarking of foundation models and classical baselines for microbiome-based disease prediction," *Research Square*, preprint, doi:10.21203/rs.3.rs-8912605/v1, 2026.

[25] J. Xing, H. Wang, Y. Sun, X. Su, and S. Wu, "Harnessing large language models to advance microbiome research: From sequence analysis to clinical applications," *Advanced Intelligent Discovery*, p. e202500038, 2025.

[26] B. Yan et al., "Recent advances in deep learning and language models for studying the microbiome," *Front. Genet.*, vol. 15, 2025, doi:10.3389/fgene.2024.1494474.

[27] C. K. Park et al., "Development of large language model specialized into microbiome datasets," *J. Microbiol. Biotechnol.*, vol. 36, p. e2511050, 2026.

[28] J. Cao, X. Xu, and X. Zhang, "Application of machine learning and large language model module for analyzing gut microbiota data," in *Advanced Intelligent Computing in Bioinformatics*, Singapore: Springer, 2024, pp. 37–48.

[29] H. El Massari, N. Gherabi, F. Qanouni, and S. Mhammedi, "Diabetes prediction using machine learning with feature engineering and hyper-parameter tuning," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, 2024, doi:10.14569/IJACSA.2024.0150818.

[30] S. Tomar, D. Dembla, and Y. Chaba, "Analysis and enhancement of prediction of cardiovascular disease diagnosis using machine learning models SVM, SGD, and XGBoost," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, 2024, doi:10.14569/IJACSA.2024.0150449.

[31] H. Almalki, A. O. Khadidos, and N. Alhebaishi, "Enhancing Alzheimer's detection: Leveraging ADNI data and large language models for high-accuracy diagnosis," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, 2024, doi:10.14569/IJACSA.2024.01511134.

[32] S. Li and X. Sun, "Dialogue-based disease diagnosis using hierarchical reinforcement learning with multi-expert feedback," *(IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, 2025, doi:10.14569/IJACSA.2025.0160232.

A Lightweight and Robust APBT–LBP Deep Feature Zero-Watermarking Framework with DNA Encryption for Medical Images

Ranjan Kumar Senapati[1], Prasanth Mankar[2], B Padmaja[3],
Chilamakuru Nagesh[4], Pradeep Kumar*[5]*, Gandikota Ramu[6], Gandharba Swain[7]
Department of ECE, VNR VJIET, Hyderabad, Telangana, India, 500018[1]
Department of ECE, Vasavi College of Engineering, Ibrahimbagh, Hyderabad, Telangana, India, 500031[2]
Department of CSE (AI & ML), Institute of Aeronautical Engineering, Hyderabad, Telangana, India, 500043[3]
Department of CSE, Srinivasa Ramanujan Institute of Technology, Anantapur, Andhra Pradesh, India, 515405[4]
Department of ECE, VNR VJIET, Hyderabad, Telangana, India, 500018[5]
Department of CSE, Aditya University, Surampalem, Andhra Pradesh, 533437, India[6]
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522302, Guntur, Andhra Pradesh, India[7]

***Abstract*—The protection of medical images in healthcare services has become important due to the adoption of telemedicine and cloud-based healthcare services. Conventional watermarking methods embed information directly into the host image, which may introduce subtle distortions and potentially affect diagnostic accuracy. This paper addresses the problem by introducing a robust zero-watermarking scheme that preserves the original medical image without any modification. The proposed approach is designed on a hybrid feature extraction scheme. Initially, desired regions of interest are identified using variance-based analysis. Local Binary Pattern (LBP) is then applied to capture fine texture details. Further, the All-Phase Biorthogonal Transform (APBT) is used to obtain stable low-frequency information. These features are subsequently pipelined through the lightweight VGG16 convolutional neural network to extract high-level semantic representations. The resulting features are fused, normalized, encrypted, and converted into binary form using Quantization Index Modulation (QIM). DNA encryption is applied to the watermark to produce a secure zero-watermark key that is stored externally. Extensive experiments conducted under a wide range of signal processing and geometric attacks show the effectiveness of the proposed method. The results show high normalized correlation above 0.99, bit error rates ($\leq 0.1\%$), and complete preservation of higher image quality, making the approach suitable for medical image authentication and copyright protection.**

***Keywords*—Zero watermarking; medical image authentication; deep learning; APBT transform; local binary pattern; CNN feature extraction; robust watermarking**

I. Introduction

The rapid growth of telemedicine and cloud-enabled healthcare platforms has significantly changed how medical images are stored, accessed, and exchanged. Today, clinicians can review patient data and collaborate with specialists across different locations almost instantly, which has greatly improved the speed and quality of diagnosis and treatment. At the same time, this increased digital connectivity brings new challenges. Medical data is now more vulnerable to unauthorized access, manipulation, and misuse. Since medical images are directly involved in clinical decisions, ensuring their authenticity and rightful ownership is critical. Even a small modification in an image can result in incorrect diagnosis, potentially affecting patient safety and raising serious legal concerns.

Conventional watermarking methods secure digital images by embedding ownership information within the host image. These approaches work well for general multimedia data. However, such an approach is not always appropriate for medical images [1], [2], [3], [4]. This is due to the fact that any modification, even if it is not perceived by the human eye, can influence diagnostic evaluation. To circumvent this deficiency, zero-watermarking has gained attention as a safe side alternative. It derives a watermark from the inherent characteristics of the image. Instead of altering the original image. Hence, preserve its quality entirely.

However, the effectiveness of zero-watermarking strongly depends on how reliably image features can be retrieved. In a practical scenario, medical images often undergo operations such as signal processing and geometric transformations. These processes can distort image features and, hence, make it difficult to regenerate a good-quality watermark. Therefore, developing feature extraction mechanisms that are stable and distinctive under attack conditions remains an important research challenge.

Among various feature descriptors, texture-based methods like Local Binary Pattern (LBP) have shown good performance in capturing fine local details [5]. LBP is simple to compute and relatively robust against illumination and noise. This makes it suitable for medical imaging applications. It focuses on local patterns, which means it may overlook the overall structure of the image. To address this gap, it is important to combine LBP with other techniques that can capture more global and stable information.

Transform-domain methods offer an additional advantage by representing image content in terms of frequency components, where low-frequency information is generally more resistant to common image processing operations. At the same time, deep learning models have demonstrated remarkable

*Corresponding author.

capability in extracting meaningful and highly discriminative features. In particular, convolutional neural networks such as VGG16 can learn complex image representations that traditional methods may fail to capture. However, deep features alone may not always remain consistent under severe distortions.

Motivated by these observations, this work introduces a hybrid zero-watermarking framework that integrates multiple feature extraction strategies. Initially, stable regions of interest (ROI) are selected based on variance analysis, ensuring that features are derived from reliable portions of the image. LBP is then used to capture local texture characteristics, while the same regions are processed using the All-Phase Biorthogonal Transform (APBT) to obtain robust low-frequency information. These combined features are further refined using the VGG16 network to extract higher-level representations. By merging handcrafted and deep features, the proposed approach aims to improve both robustness and consistency.

Finally, the extracted features are used to generate a secure zero-watermark key, which enables ownership verification without modifying the original image. Through this multi-layered feature integration, the proposed method attempts to overcome the limitations of existing techniques and offers a more reliable solution for protecting medical image ownership in modern healthcare environments.

The major contributions of this work are as follows:

- A hybrid deep learning-based zero-watermarking framework is proposed for medical image authentication without modifying the original image.
- Stable ROI blocks are selected using variance analysis to ensure robust feature extraction.
- LBP, APBT, and CNN (VGG16) are combined to extract complementary texture, frequency, and deep features.
- QIM is used to binarize features into a compact and discriminative sequence.
- ORB-based alignment is applied to handle geometric distortions.
- Robustness is validated against 42 signal processing and geometric attacks.

II. Related Work

Conventional feature extraction algorithms such as SIFT, SURF, and KAZE, are often combined with transformed domain watermarking to achieve robustness to rotation, scale, and translation asserts. Binary feature extraction algorithms such as BRIEF, ORB, and BRISK have faster run time, making them suitable for real-time applications. These algorithms also combine with transformed domain watermarking techniques. Several works have been reported with these combinations.

In this context, Senapati et al. [6] combined the SIFT feature with DTT to improve the robustness to RST attacks. Similarly, Gao and Chen [7] applied a fusion technique, which includes DWT-SVD with SURF, RANSAC and improved ABC algorithms to improve security, robustness, and imperceptibility of the watermark scheme. Further, Soualmi et al. [8] proposed a medical image watermarking approach for tamper detection that combines SURF keypoints, Weber descriptors, and the Arnold transform. The method identifies salient points within the ROI using SURF and embeds the watermark in their surrounding regions, while Weber descriptors are used for reliable watermark embedding and extraction. Another work carried by Ray et al. [9] utilizes KAZE features to construct a perceptual hash of the watermark, which is then embedded using transform-domain techniques (LWT-SVD). The approach improves robustness and flexibility by removing size constraints on the watermark and enhancing resistance to attacks.

In contrast to the above methods, which embed watermark information into the host image, zero-watermarking methods generate a watermark signature based on extracted image features. Several studies have used moment invariants, texture descriptors, and transform-domain coefficients for zero watermark generation [10], [11]. Orthogonal moments zero-watermarking, such as Zernike [12], Tchebichef [13], Legendre-Fourier [14], etc. have been proposed because these are inherently invariant to scaling, translation, and rotation. These approaches construct a feature matrix by computing significant moment parameters.

Based upon these moment-based techniques, a multiple-zero-watermarking framework for protecting medical images in IoMT applications is reported in [15]. It uses multi-channel fractional Legendre Fourier moments (MFrLFM) due to its high stability, accuracy, and geometric invariance. The scheme is realized on a Raspberry Pi board to evaluate its aptness for IoMT applications. However, the method is evaluated under a few attack scenarios.

To further enhance robustness and resolve the computational issues, a zero-watermarking scheme based on fractional-order radial harmonic Fourier moments (FoRHFMs) is proposed by [16]. The authors tackle the issue of numerical instability, which is inherent in continuous orthogonal moments, by proposing integer-order radial harmonic Fourier moments (IoRHFMs) to their fractional-order counterparts (FoRHFMs). This improves the computational stability and reliability. The experimental results show that this algorithm is resistant to lossless copyright protection of medical images and robust to all kinds of attacks.

In addition to moment-based approaches, hybrid feature-based schemes have also been investigated. Gharib et al. [17] presented a zero-watermarking method for color images. LBP and VGG-19 features are fused to generate a feature matrix. DWT and DCT improve the feature representation. The encrypted watermark is XORed with the improved feature matrix to generate a zero-watermark. Higher robustness and imperceptibility are achieved for a set of 12 images. However, this method has no stable ROI, and geometric alignment robustness is missing. In contrast to zero-watermarking approaches, embedding-based methods are still being investigated for improved performance. Anand et al. [18] proposed a multimodality image fusion scheme using NSST and DTCWT to generate a watermark image. The combination of NSST, BEMD, and MSVD conceals the generated mark in the host image. Weighted HOG-based optimization is applied to trade off between imperceptibility and robustness. The method shows an improvement of 47.65% compared with a standard

baseline scheme. Unlike [18], which embeds a watermark using NSST–BEMD–MSVD, our proposed method avoids modifying the host image and ensures higher medical data integrity.

To overcome the limitations of handcrafted and transform-based features, recently, deep learning-based watermarking has become popular among researchers. This is due to the fact that many pretrained models can be directly used for extracting the deep semantic features. The encrypted watermark can be embedded with the deep features. Further, deep learning can adapt to different image features to enhance higher robustness and imperceptibility. However, the unique nature of medical imaging data poses significant challenges, including limited availability of large-scale annotated datasets and the high cost associated with expert labeling. These constraints make it difficult to train deep neural networks from scratch for reliable feature extraction. These limitations are circumvented using transfer learning, where knowledge from a pre-trained model is used. The pre-trained model is trained on a large dataset and adapted to a target domain with limited data. This improves generalization performance and reduces training requirements.

Based on these advancements, several deep learning-based watermarking algorithms have been proposed. Fan et al. [19] extract image features using Inception V3 CNN and then encrypt the hidden watermark using the chaotic map. The method shows robustness to a wide range of geometric assertions. However, it is less robust to common attacks. Similarly, Dong et al. [20] used an enhanced NasNet-mobile CNN. It also used DCT features to embed watermarks. The scheme shows strong resistance to most of the assaults. In addition, a multiwatermarking algorithm approach using GoogleNet transfer learning for medical images is proposed by Zhang et al. [21]. Henon chaos encryption is used to hide multiple pieces of watermark information. The deep semantic features XORed with the secret watermark, and the generated key is stored in a 3rd party authority. The framework provides good robustness to conventional and geometrical attacks. Similarly, Sheng et al. [22] used ResNet50 to extract deep features. Then DCT transformation and a perceptual hash function are used to generate a feature vector.

Recently, a Swin Transformer-based zero watermarking is suggested by Han et al. [23] to improve the security of medical images in the metaverse healthcare system. This Swin Transformer is pretrained. The deep features generated from the host medical images by the transformer show good generalization performance. A mean hashing algorithm is used to generate multiscale and binary feature vectors. The watermark is encrypted using a logistic chaotic encryption method, and it is XORed with the binary feature vector to create a zero-watermark. The scheme exhibits robustness against all kinds of attacks and provides privacy protection for medical image transmission in the metaverse.

In a related direction, another work by Han et al. [24] addresses the critical issue of protecting privacy and ensuring data security in teledermatology systems within the IoMT. The zero-watermarking method uses federated learning and sparse autoencoders to extract secure image features without altering the original medical image. By applying 2D-DCT on these features, the scheme creates a watermark that is robust to attacks while preserving image integrity and patient privacy.

Furthermore, a zero-watermarking algorithm using multiple circular statistical features of the medical image is proposed by Wang et al. [25]. It combines DCT and SVD for feature stability and uses Arnold transform with XOR operation to securely generate the zero watermark without modifying the original image. The method also includes a blind image correction technique and demonstrates strong robustness against noise, filtering, and geometric attacks.

Motivated by the above limitations and improvements, the proposed zero-watermarking framework, extracting stable and discriminative features from medical images is a critical step. Initially, a stable Region of Interest (ROI) is identified to ensure robustness against most of the attacks. The selected ROI is then transformed using the All-Phase Biorthogonal Transform (APBT) to obtain invariant frequency-domain representations. These transformed coefficients are subsequently fed into a pretrained VGG16 model, originally trained on the large-scale ImageNet dataset, to extract deep hierarchical features. To further adapt the model to the medical domain, fine-tuning is performed using the constructed medical image dataset, enabling the network to capture domain-specific characteristics more effectively. The resulting deep features are then quantized using Quantization Index Modulation (QIM) to generate a robust binary sequence, which serves as the zero watermark. This integration of ROI selection, APBT transformation, transfer learning-based deep feature extraction, and QIM-based encoding ensures improved resistance and stability of the proposed zero-watermarking scheme under various attacks.

III. Preliminaries

A. All-Phase Biorthogonal Transform

All-phase Biorthogonal Transform (APBT) can be understood as a sequency-filtered version of the DCT. The idea is that APBT modifies the conventional DCT basis through all-phase preprocessing and sequency filtering, which improves phase consistency and energy compaction [26]. The APBT has better energy compaction in low frequencies and attenuation characteristics at high frequencies.

Suppose χ is an image which is cropped into $B \times B$ image blocks. and A represents the APBT matrix of the block. The APBT transform coefficients 'Y' can be expressed as:

$$Y = A\chi A^T \tag{1}$$

where, A^T is the transpose matrix of χ. To reconstruct the image χ, we can write

$$\chi = A^{-1}Y(A^{-1})^T \tag{2}$$

A^{-1} is the inverse matrix of A.

The transform matrix A is set in different forms as specified in Eq. 3 and 4. The corresponding transforms are denoted as an APBT and its inverse, IAPBT, respectively.

$$A_1(p,q) = \begin{cases} \frac{B-p}{B^2}, & p = 0,1,..B-1, q = 0 \\ \frac{(B-p)cos\frac{pq\pi}{B} - csc\frac{q\pi}{B} sin\frac{pq\pi}{B}}{B^2}, & p = 1,2,....B-1, \\ & q = 1,2,....B-1. \end{cases} \tag{3}$$

$$A_2(p,q) = \begin{cases} \frac{1}{B}, & p = 0, q = 0,1,..B-1. \\ \frac{B-p+\sqrt{(2)}-1}{B^2}.cos\frac{p(2q+1)\pi}{B}, & p = 0,1,2.....B-1, \\ ; & q = 1,2,....B-1. \end{cases} \tag{4}$$

B. DNA Encryption and Decryption

DNA encryption is employed on VGG features and the watermark to enhance the security of the zero-watermark. The binary watermark is first encoded into a DNA chain using a predefined encoding algorithm coupled with logical operations such as complement and permutation. This strategy significantly increases the key space and ensures strong confusion and diffusion operation. This enhances the system's ability to prevent brute-force attacks. A brief description of the DNA encryption and decryption mechanism is as follows:

DNA is composed of four fundamental nucleic acid bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). These bases store genetic information in the form of coded sequences. According to the Watson–Crick pairing principle, Adenine pairs with Thymine (A–T) and Guanine pairs with Cytosine (G–C) [27].

In DNA-based data representation, binary information can be mapped to these bases by encoding every two bits into one nucleotide. Based on the encoding and decoding constraints governed by complementary pairing rules, eight valid encoding schemes can be defined, as illustrated in Table I.

TABLE I. DNA ENCODING RULE

Rules	Watson-Crick Complementary Rules							
	1	2	3	4	5	6	7	8
00	A	A	C	G	C	G	T	T
01	C	G	A	A	T	T	C	G
10	G	C	T	T	A	A	G	C
11	T	T	G	C	G	C	A	A

In addition to encoding and decoding, DNA sequences can perform several operations such as addition, subtraction, complementation, and XOR. Nevertheless, the XOR operation is widely used in encryption for its simplicity and reversibility. Table II shows the DNA-based XOR operation, where each nucleotide in a row is combined with a nucleotide in a column to produce a corresponding output.

TABLE II. DNA XOR OPERATION BETWEEN EACH ELEMENT IN ROW AND EACH ELEMENT IN COLUMN

XOR row/column	A	C	G	T
A	A	C	G	T
C	C	A	C	G
G	G	T	A	C
T	T	G	T	A

To illustrate the encoding and encryption process, consider a binary sequence "10101101". This sequence is first divided into 2-bit groups and encoded into DNA bases according to a selected rule from Table I. For example, using Rule 2, the sequence is converted into a DNA string. Next, a predefined DNA key (e.g., "ACGT") is applied, and a DNA XOR operation is performed between the encoded sequence and the key, resulting in a new DNA sequence. This sequence is then decoded back into a binary form using the same encoding rule, producing the encrypted binary output.

The decryption process follows the reverse steps, where the encrypted binary sequence is converted back to DNA bases, XORed with the same key, and finally decoded to recover the original binary data. The pseudocode for the DNA-based encryption process is provided in Algorithm 1.

Unlike conventional encryption methods such as AES and RSA, DNA provides biologically inspired encoding. The XOR-based operation is computationally simple and well-suited for binary feature encryption. It enhances the features' confusion and diffusion properties while maintaining a lightweight zero-watermark operation. In the proposed scheme, DNA encryption is applied only to compact binary features instead of the full image. The objective of selecting DNA scheme is not to replace other cryptographic algorithms, but to provide a lightweight security layer for processing in real-time.

Algorithm 1 DNA-Based Feature Encryption

Require: Binary feature vector F_b
Ensure: Encrypted DNA sequence E_{dna}
1: Convert binary vector F_b into DNA sequence F_{dna} using encoding rule R
2: Generate random DNA key sequence K_{dna} of the same length as F_{dna}
3: **for** $i = 1$ to length(F_{dna}) **do**
4: $E_{dna}(i) \leftarrow$DNA_XOR($F_{dna}(i), K_{dna}(i)$)
5: **end for**
6: **return** E_{dna}

C. Feature Extraction

Efficient feature extraction is a vital step to ensure resilience under various attacks. We have proposed two complementary methods to extract features from the host medical image. LBP extracts texture features from the image, while VGG16 extracts deep-level semantic features.

1) Local binary pattern: Local Binary Pattern (LBP) is known as a texture descriptor. It efficiently encodes the local structure of an image by comparing a pixel intensity with the neighboring pixel intensities [28]. The image is partitioned, and LBP values are estimated and finally combined into a global representation. The computation of the LBP value for a given pixel is as follows:

$$S(p,q) = \begin{cases} 1, \text{if } \ g(p) \geq g(q) \\ 0, \text{otherwise} \end{cases} \tag{5}$$

The $g(p)$ and $g(q)$ in Eq. 5 represent the pixel intensities of the neighboring and center pixel, respectively. This operation provides texture analysis while ensuring robustness across various illumination conditions. Fig. 1 shows the generated texture map from the host medical image.

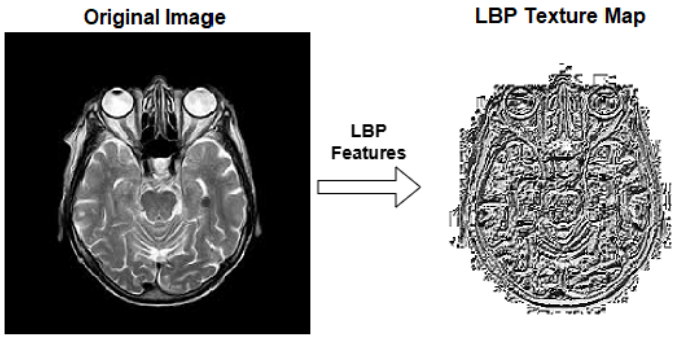


Fig. 1. LBP feature map of the "brain" image, showing pixel-wise texture encoding.

2) VGG-16 deep features: The VGG16 is a deep CNN that consists of 13 convolution layers and 2 fully connected (FC) layers. This gives a total of 15 layers. The output of the final FC is connected to a ReLU layer, which maps all the features to be positive. The VGG16 is a pretrained CNN that has weights and biases trained by ImageNet. It generates hierarchical features. The output feature map of the layer $X^{(l)}$ after convolution, activation, and pooling is fed to the next layer. In VGG16, the deep feature vector is obtained after the final convolution block and flattening.

The equation for deep features:

$$Z^{(l)} = W^{(l)} * X^{(l-1)} + b^{(l)} \tag{6}$$

$$A^{(l)} = \text{ReLU}\left(Z^{(l)}\right) \tag{7}$$

$$P^{(l)} = \text{MaxPool}\left(A^{(l)}\right) \tag{8}$$

$$X^{(l)} = P^{(l)} \tag{9}$$

$$f = \text{vec}\left(X^{(L)}\right) \tag{10}$$

$$F = \sigma(Wf + b) \tag{11}$$

where,

- $X^{(l-1)}$: Input feature map of layer $(l-1)$
- $W^{(l)}$: Convolution kernel weights
- $b^{(l)}$: Bias term
- $*$: convolution operation
- $\sigma(.)$: Nonlinear activation function (ReLU)
- $P^{(l)}$ Pooled feature map
- $X^{(l)}$: Output feature map of layer
- $X^{(L)}$: Denotes the output feature map from the final convolutional layer.
- f: Flattening operation,
- W: Contains the learned parameters that connect the input feature vector to the output neurons. These values are learned during the training phase of the CNN.
- F: Fxtracted deep feature descriptor used for watermark generation

$$f \in \mathbb{R}^{25088}, \text{and } F \in \mathbb{R}^{4096} \tag{12}$$

The number 25088 represents the total number of elements in the final convolution feature map of VGG16 before it enters the fully connected layers. This vector is passed to the fully connected layer (fc6) and then fc7, which are designed with 4096 neurons. fc7 output passes through ReLU, and the resulting 4096-dimensional vector is typically used as the deep feature descriptor.

Fig. 2 shows the VGG16 network used in the proposed method to extract deep semantic features at the output of the ReLU layer. The network learn increasing abstract visual representation at the output of each layer

Feature Fusion: The APBT, LBP, and VGG16 descriptors are integrated to form a unified feature representation, improving the discriminative capability of the system. LBP effectively characterizes local texture patterns, whereas VGG19 extracts high-level semantic and structural information from the image. The fused feature vector captures both regional texture details and global contextual information. This fused representation improves the robustness of the model against a wide range of image attacks.

IV. Proposed Zero Watermarking Framework

The proposed zero-watermarking system is depicted in Fig. 3. It consists of two major stages: registration and verification.

In the registration stage, the APBT and LBP feature sets are encrypted using a DNA algorithm before being fed to the VGG model for high-level semantic representation. After feature normalization and binarization, the watermark is encrypted separately using the same DNA encoding scheme. A hash derived from the host image is subsequently incorporated to bind the watermark to the host medical image. Finally, a zero-watermark key is generated and securely stored without altering the original image. Algorithm 2 presents the proposed zero-watermarking method.

The registration pipeline includes:

- Input medical image followed by preprocessing
- Stable ROI selection
- APBT-LBP feature extraction and Feature fusion
- DNA encryption
- CNN-based deep feature extraction
- Feature binarization using QIM
- Zero watermark generation
- Key storage
- Watermark verification

In the verification stage, the distorted test image is aligned with the original using ORB feature-based matching to handle

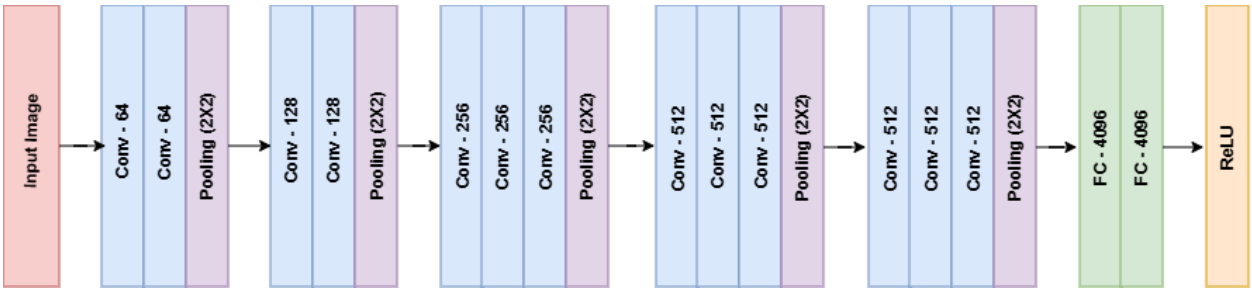


Fig. 2. VGG16 architecture used for feature map extraction.

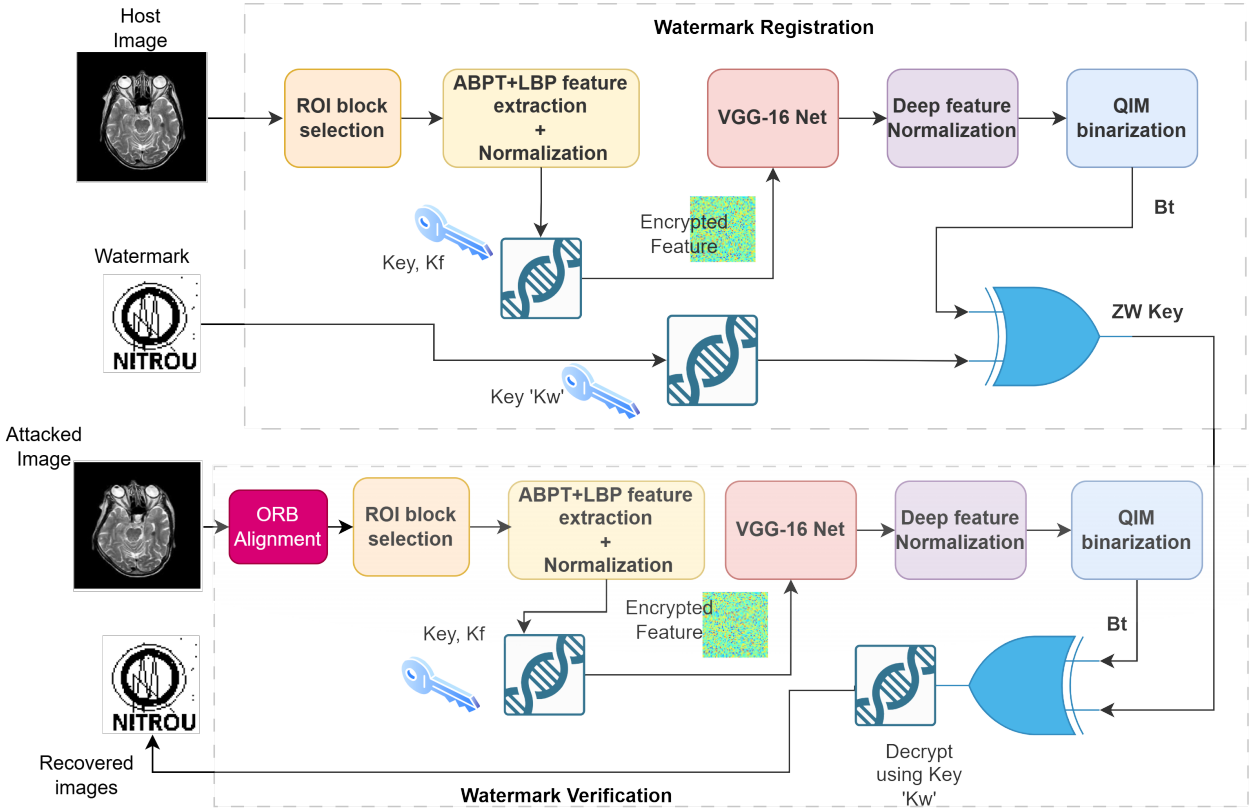


Fig. 3. Block diagram of the proposed scheme.

geometric imbalance. The same stable regions are then used to match feature consistency. The regions undergo the encoding, deep feature extraction, normalization, and binarization. The encrypted watermark is retrieved and decrypted using the stored zero-watermark key. Finally, the retrieved watermark is compared with the original watermark to evaluate robustness. Algorithm 3 presents the proposed watermark verification method.

The verification pipeline includes:

- Input attacked image
- Geometric alignment
- stable ROI extraction
- APBT+LBP Feature extraction
- Deep feature generation
- Binary feature generation using QIM
- Watermark recovery
- DNA decryption
- Recovered watermark

A. Stable ROI Block Selection

The input medical image is divided into non-overlapping blocks of size 8×8. For each block, the variance is computed:

$$Var(B) = \frac{1}{N} \sum_{i=0}^{N} (B_i - \mu)^2 \tag{13}$$

Blocks with variance greater than a predefined threshold are considered stable blocks. These blocks contain rich struc-

tural information and are less sensitive to image processing attacks.

B. APBT+LBP Feature Extraction

After the selection of stable ROI blocks based on variance, APBT, and LBP features of the ROI blocks are concatenated. APBT is effective in preserving signal energy and suppressing boundary artifacts. Only the low-frequency components of size (4×4) are retained in APBT because they are less sensitive to common signal processing operations and contribute to the stability of the feature representation against various attacks. LBP encodes the relationship between a pixel and its neighboring pixel to generate a compact representation of micro-patterns such as edges and textures. This enhances the robustness against illumination changes and noise variations. A combination of these features enables the algorithm to extract robust features. The features are then normalized and concatenated to form a unified feature vector. The aggregation of these features results in a comprehensive feature map, which is subsequently used for further processing in the zero-watermark generation. The details of the mathematical steps are developed below.

The low-frequency coefficients of APBT for each selected ROI block can be represented as:

$$F_{APBT} = \{c_1, c_2, \ldots, c_m\} \tag{14}$$

where, c_i denotes the normalized low-frequency APBT coefficients of the selected ROI blocks. The subsequent LBP texture features are computed from the same ROI regions and represented as

$$F_{LBP} = \{l_1, l_2, \ldots, l_n\} \tag{15}$$

where, l_i represents the normalized LBP histogram features.

The fusion of transform-domain and texture-domain features are concatenated to form a discriminative hybrid feature vector:

$$F = [F_{APBT}; F_{LBP}] \tag{16}$$

The fused feature vector is reshaped into a feature image and resized to 224×224 to be compatible with the pretrained VGG-16 network. Deep features are subsequently extracted from the fully connected layer ($fc7$) of VGG16 and expressed as:

$$F_d = \phi_{VGG16}(F) \tag{17}$$

The $\phi_{VGG16}(\cdot)$ denotes the nonlinear deep feature mapping function of the pretrained VGG16 model. Since the network parameters remain fixed during execution, the possibility of overfitting is reduced. The combination of ABPT-LBP and deep features improves feature generalization. In the subsequent operations, the features are normalized prior to QIM-based binarization and watermark generation.

C. Feature Normalization

The feature vector is normalized using z-score normalization:

$$F_n = \frac{F_d - \mu}{\sigma} \tag{18}$$

where, μ and σ represent the mean and standard deviation of the features.

D. Feature Binarization Using QIM

Quantization Index Modulation is used to convert deep features into binary form:

$$b_i = mod(\lfloor F_n/\Delta \rfloor, 2) \tag{19}$$

where, Δ is the quantization size.

E. Zero Watermark Generation

Let B = binary feature vector and ϑ = binary watermark. Then the zero watermark key is generated as

$$ZW = B \oplus \vartheta \tag{20}$$

The generated key and the associated encryption parameters are stored externally. In practical deployment, the key is securely maintained within hospital Picture Archiving and Communication System (PACS) servers, cloud storage systems, or blockchain-assisted authentication frameworks. In the proposed method, the extracted feature vector and watermark information are encrypted using DNA-XOR operations with random Keys K_f and K_w respectively. This makes the unauthorized reconstruction highly cumbersome. Moreover, the stored zero-watermark data contains only an encrypted bit string, not the image. This reduces the risk of sensitive medical data leakage.

F. Watermark Verification

During verification, the same feature extraction pipeline is applied to the test image. To compensate for geometric distortions, ORB-based feature matching and geometric transformation estimation are performed before feature extraction.

The recovered watermark is obtained as:

$$\vartheta_{enc} = B_t \oplus ZW_{key} \tag{21}$$

where, B_t is the binary feature vector extracted from the test image.

V. Experimental Setup

The experiment was carried out using MATLAB 2024a on a system equipped with an Intel i5 processor and 16GB of RAM. We have used different medical images of different sizes as host medical images (HMIs). These include COVID-19 chest X-ray sizes 3050×2539, 1698×1878, 1064×840 [29], MRI brain images of sizes 350×350, 210×240, 377×500, CT images of size 512×512 [30].

Algorithm 2 Proposed Zero Watermarking Framework

Require: Host medical image χ, watermark ϑ
Ensure: Zero-watermark key ZW_{key}
Resize χ to 256×256 and divide into 8×8 blocks
for each block $B_{i,j}$ **do**
  Compute variance $\sigma^2(B_{i,j})$
  if $\sigma^2(B_{i,j}) > T$ **then**
    Select block as Stable ROI
    Apply APBT transform
    Extract low-frequency coefficients
    Normalize coefficients
    Extract LBP texture features
    Normalize LBP vector
    Fuse features:

$$F = [F_{APBT}; F_{LBP}]$$

  end if
end for
Construct feature vector F_{map}
Convert feature vector to binary
Apply DNA encoding
Generate random DNA key K_f
Encrypt features using DNA-XOR
Convert encrypted features back to decimal
Reshape encrypted feature vector to image F_{img}
Resize F_{img} to 224×224 and replicate to 3 channels
Extract deep features using VGG16 layer $fc7$
Normalize deep features:

$$F_n = \frac{F - \mu}{\sigma}$$

Apply QIM binarization:

$$B = mod(\lfloor F_n/\Delta \rfloor, 2)$$

Convert watermark ϑ to binary vector
Apply DNA encoding to watermark
Generate random DNA key K_w
Encrypt watermark using DNA-XOR
Convert encrypted watermark back to binary
Generate zero watermark key:

$$ZW_{key} = B \oplus \vartheta_{enc}$$

Store $\{ZW_{key}, K_w, K_f, \mu, \sigma, \Delta\}$

Algorithm 3 Zero Watermark Verification

Require: Test image κ, stored data
Ensure: Recovered watermark ϑ_r
Apply attacks to κ
Align test image using ORB feature matching
Extract Stable ROI blocks using stored coordinates
for each selected block **do**
  Extract APBT + LBP fused features
end for
Construct feature vector F_t
Apply same DNA encryption using key K_f
Convert encrypted vector to feature image
Extract deep features using VGG16 ($fc7$)
Normalize using stored parameters
Apply QIM binarization to obtain B_t
Recover encrypted watermark:

$$\vartheta_{enc} = B_t \oplus ZW_{key}$$

Apply DNA decryption using key K_w
Convert binary to watermark image
Output recovered watermark ϑ_r

The COVID-19 chest X-ray images used in this study were originally acquired at high spatial resolution. Before embedding, the host images were resized to 224×224 pixels to ensure compatibility with the pretrained VGG-16 and to maintain computational efficiency. Resizing COVID-19 images introduces significant spatial downsampling. The proposed framework is designed for zero-watermark generation, which extracts stable and discriminative features rather than clinical diagnosis or lesion detection. Moreover, our scheme does not modify the original medical image, preserving its diagnostic integrity in practical deployment. The resizing operation primarily affects the feature extraction stage. It does not alter the actual medical image stored or used by doctors for diagnosis. Future work may investigate multi-scale feature extraction frameworks as a preprocessing stage to better preserve subtle pathological details.

A copyrighted image of size $64 \times 64$was employed. These HMIs and corresponding watermark images are illustrated in Fig. 4. To assess the visual quality of the watermarked images, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were used. The algorithm's robustness was measured using metrics such as Normalized Correlation Coefficient (NCC) and Bit Error Rate (BER). PSNR and SSIM reflect the perceptual similarity between the original and watermarked images, while NCC and BER assess the strength and reliability of the embedded watermark. Ideally, NCC should be close to 1, indicating high similarity, and BER should be 0, reflecting superior robustness. The mathematical definitions of these evaluation metrics are presented below.

$$\text{PSNR}(\chi, \kappa) = 10log_{10}\frac{\chi^2_{peak}}{\Psi}, \tag{22}$$

where, Ψ is the mean square error between the test and host medical image. Evaluating image degradation through SSIM involves perceiving changes in structural information and considering perceptual phenomena such as luminance and contrast masking. SSIM addresses the limitations of PSNR and is expressed in Eq. (23) by:

$$\text{SSIM}(\chi, \kappa) = \frac{(2\mu_\chi\mu_\kappa + a_1) + (2\sigma_{\chi\kappa} + a_2)}{(\mu^2_\chi + \mu^2_\kappa + a_1)(\sigma^2_\chi + \sigma^2_\kappa + a_2)}, \tag{23}$$

where, μ_χ, μ_κ and σ_χ, σ_κ are the mean and variance of image χ and κ respectively. $\sigma_{\chi\kappa}$ is the covariance between χ and κ. a_1 and a_2 are +ve constants. For an watermark size $\vartheta = N_1 \times N_2$, the NCC between the ϑ and ϑ_r can be represented by [31]:

$$\text{NCC}(\vartheta, \vartheta_r) = \frac{\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} \vartheta(n_1,n_2) \times \vartheta_r(n_1,n_2)}{\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} \vartheta(n_1,n_2)^2}. \quad (24)$$

Similarly, the BER defines the percentage of error between ϑ and ϑ^*, which can be expressed by:

$$\text{BER}(\vartheta, \vartheta_r) = \frac{1}{S} \times \sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} [\vartheta(n_1,n_2) - \vartheta_r(n_1,n_2)] \quad (25)$$

VI. Results and Discussion

This section presents the robustness evaluation of the proposed scheme along with a comparative analysis against existing methods. Finally, the Wilcoxon signed-rank statistical test and the Friedman ranking test are conducted to evaluate the effectiveness of our method.

A. Robustness Evaluation

The robustness of the proposed framework was tested against 42 different attacks, including: Noise Attacks, Gaussian noise, Salt and pepper noise, Speckle noise, Filtering Attacks, Gaussian filtering, Median filtering, Average filtering, Compression Attacks, JPEG compression at different quality levels, Geometric Attacks: Rotation, Scaling, Translation, Cropping, Intensity Attacks, Histogram equalization, Gamma correction, Contrast stretching, Advanced Attacks: Motion blur, Poisson noise, Elastic distortion, Patch removal, Cut-and-paste manipulation, The watermark recovery performance remains high even under severe distortions.

Fig. 5 shows the various signal processing attacks applied to the test image and the corresponding recovered watermarks. The numerical values of NCC and BER are also listed under the recovered watermarks. The applied sets of combined attacks are 1) No attack 2) Average filtering (3x3), 3) Contrast stretch, 4) Gamma correction by (0.5 and 1.5), 5) Gaussian (0.01), 6) Histogram equalization, 7) Gaussian blur (σ=1), 8) Histogram compression, 9) Sharpen, 10) Salt and Pepper (0.02), and 11) JPEG (Q=50). A very high robustness is observed in all attack scenarios. The NCC values consistently above 0.99, indicating near-perfect watermark recovery. Correspondingly, BER remains very low, ranging from 0.05-0.3%. These quantitative results confirm that the proposed scheme maintains strong resilience against diverse signal distortions, including Noise, gamma corrections, histogram manipulations, Gaussian blur, compression, and sharpening operations. The retrieved watermarks are graphically illustrated in Fig. 5 under different attacks.

Fig. 6 shows the various geometrical, combined attacks applied to the test image and the recovered watermarks. The numerical values of NCC and BER are also listed. The applied sets of combined attacks are 1) Blur+noise, 2) Rotate by 5°+blur, 3) Rotate by 5°+noise, 4) Scale+JPEG, 5) Random bending, 6) Rotate by 5°, 7) Rotate by 10°, 8) Motion blur, 9) Cut-paste, 10) Elastic distortion, 11) Patch removal, and 12) scale by 0.8. A very high robustness is observed in all attack scenarios. The NCC values consistently above 0.99, indicating near-perfect watermark recovery. Correspondingly, BER remains very low, ranging from 0.05-0.3%. These quantitative results confirm that the proposed scheme maintains strong resilience against diverse geometric distortions, including arbitrary-angle rotation, elastic distortion, and scaling operations. The retrieved watermarks are graphically illustrated in Fig. 6 under different attacks.

Robustness to geometric attacks is closely related to ROI selection and ORB-based geometric alignment. The ROI selection improves feature consistency, and ORB feature matching compensates for rotation, scaling, and geometrical distortion prior to feature extraction. The higher NCC values in Fig. 6 demonstrate that the ORB alignment stage effectively preserves feature correspondence under geometric and combined attacks.

Fig. 7 shows the NCC and SSIM performance of the proposed method against 42 attacks. It is observed that NCC and SSIM are 1 under no attack case. The NCC is very close to 1 for all attacks irrespective of common, geometrical, or combined attacks. The average SSIM is above 0.7, considering all attacks. It indicates that the algorithm shows a good trade-off between imperceptibility and robustness. The robustness performance can be shown in the BER plot as shown in Fig. 8. The BER below 0.3% indicates perfect watermark recovery.

Both Gaussian Noise, Salt & Pepper Noise have robustness between (0.98–1.0), Hoever, SSIM drops significantly (0.3–0.8). This indicates that both noise deteriorates visual quality but do not impact the quality of the watermark. NCC remains 1.0, SSIM moderately high (0.8–0.95) for Gaussian blur, median, and average filter attacks, indicating filtering preserves structural content. High robustness and visual quality indicate effective use of low-frequency and structural features. NCC and SSIM are near 1.0 across all quality factors (Q) in JPEG attacks, showing robustness against real-world compression scenarios. NCC is approximately between (0.95 to 1.0), and SSIM is slightly reduced (0.8–0.95) in most of the geometrical attacks. This indicates the geometric distortion is successfully compensated by ORB-based alignment. A Minor SSIM drop may be due to interpolation and spatial misalignment. NCC is high (0.95–1.0), but SSIM drops abruptly (i.e., 0.3–0.6) on intensity-based attacks such as histogram equalization, gamma correction, and contrast stretching. This is due to the fact that these attacks alter the pixel intensity distribution. For combined and random bending cases, NCC remains significantly high (0.95+), whereas SSIM varies widely (0.4–0.9). This indicates that watermark detection is reliable even under complex distortions.

The PSNR plot in Fig. 9 illustrates the impact of various attacks on the perceptual quality of medical images, along with deviation (error bars). It provides both average performance and stability details. The PSNR for the no-attack case is between 24 and 25 dB. This confirms that no artificial distortion is introduced by the proposed algorithm. These values show inherent dataset variability instead of embedding loss. Under Gaussian Noise and Salt & Pepper Noise, PSNR drops to a range of 21–24 dB. This indicates moderate degradation, but it is still acceptable.

PSNR remains relatively higher (25–30 dB) under Gaussian blur, median filter, and average filter. This shows that the method preserves global structures. PSNR shows up to

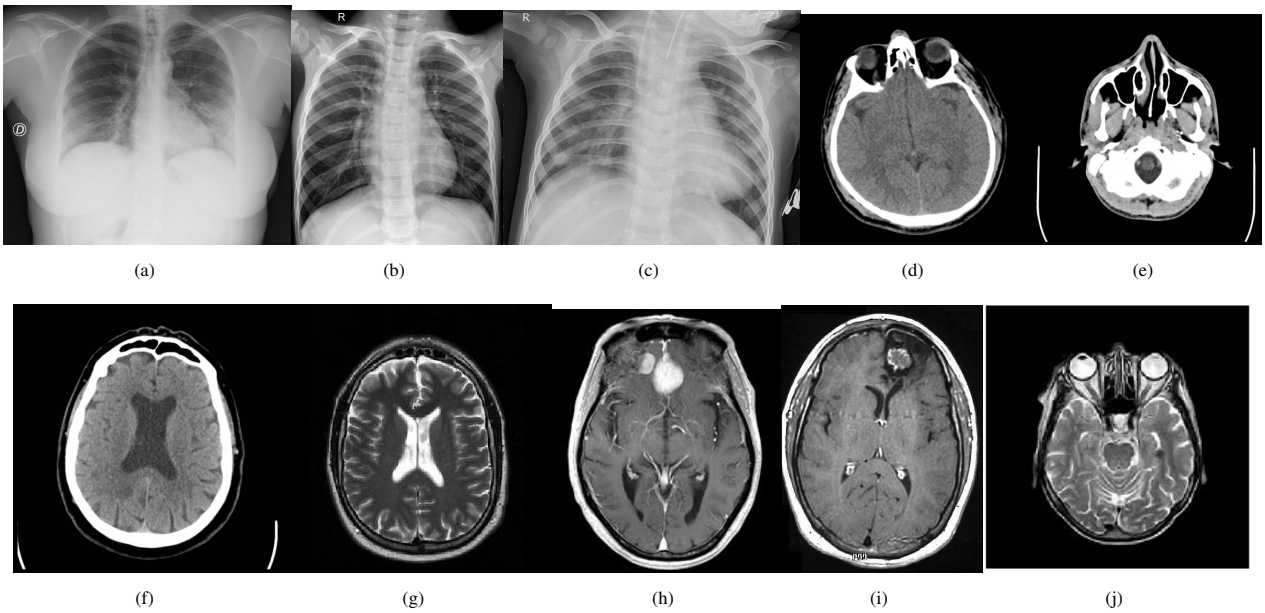


Fig. 4. Sample images used for experiment: (a), (b), (c) X-ray; (d), (e), (f) CT, and (g), (h), (i), (j) MRI images.

35–38 dB at higher quality (Q70) and drops gradually with compression strength. This indicates that compression removes redundancy but retains structural information. Therefore, the proposed algorithm leverages structure-based features. Moderate deviation in PSNR (26-34 dB) is observed in the rotation and scaling operations. This is due to ORB-based alignment. A significant PSNR drop (13–20 dB) occurs in histogram equalization and gamma correction as these attacks alter the global intensity distribution. This affects PSNR but not necessarily the watermark quality.

PSNR fluctuates widely (20–33 dB) in sharpening, motion blur, Poisson noise, and combined rotation with noise. Histogram compression/intensity scaling has the lowest PSNR (13–15 dB).

Large bar plots were observed in gamma correction, cropping, and combined attacks. This indicates performance relies on image content (edges vs smooth regions). This may be acceptable for medical images with varying textures.

The PSNR analysis demonstrates that the proposed framework preserves high perceptual quality under most signal processing and geometric attacks, while significant degradation is observed only under intensity-based transformations such as histogram equalization and gamma correction. This behavior confirms that the variations in PSNR are primarily due to attack severity rather than the watermarking process, making PSNR less indicative of robustness in zero-watermarking systems.

The performance of the proposed zero-watermarking algorithm is further analyzed using heat map visualization in Fig. 10, which provides an intuitive representation of spatial feature stability and robustness. The generated heat maps highlight the distribution of discriminative features extracted from the stable ROI using the APBT and deep features (VGG16).

From the heat map analysis, it is observed that:

- High-intensity regions correspond to structurally significant areas of the medical image, indicating that the algorithm effectively captures salient and invariant features.
- The feature concentration remains consistent across various attacks (e.g., noise, filtering, compression), demonstrating the strong robustness of the extracted watermark features.
- Minimal variation in heat distribution between the original and attacked images confirms the stability of the feature representation, which directly contributes to higher NCC values and lower BER.
- The use of stable ROI selection and ORB-based alignment ensures that the important regions retain their prominence in the heat maps even under geometric distortions.
- Compared to conventional methods, the proposed framework exhibits better localization of important features, reducing the influence of non-informative regions and improving watermark reliability.

Overall, the heat map visualization validates that the proposed method maintains spatial consistency, robustness, and discriminative feature preservation, which are critical for reliable zero-watermark generation and extraction.

B. Comparative Analysis

Table III shows the NCC comparison between the proposed method and other methods for conventional, geometrical, and combined attacks. It is observed that the proposed scheme outperforms the methods in [32] and [33] by a wide margin. Method [17] and [34] closely follows our method. The combined attack means that multiple image processing attacks are applied sequentially to the same image before extracting

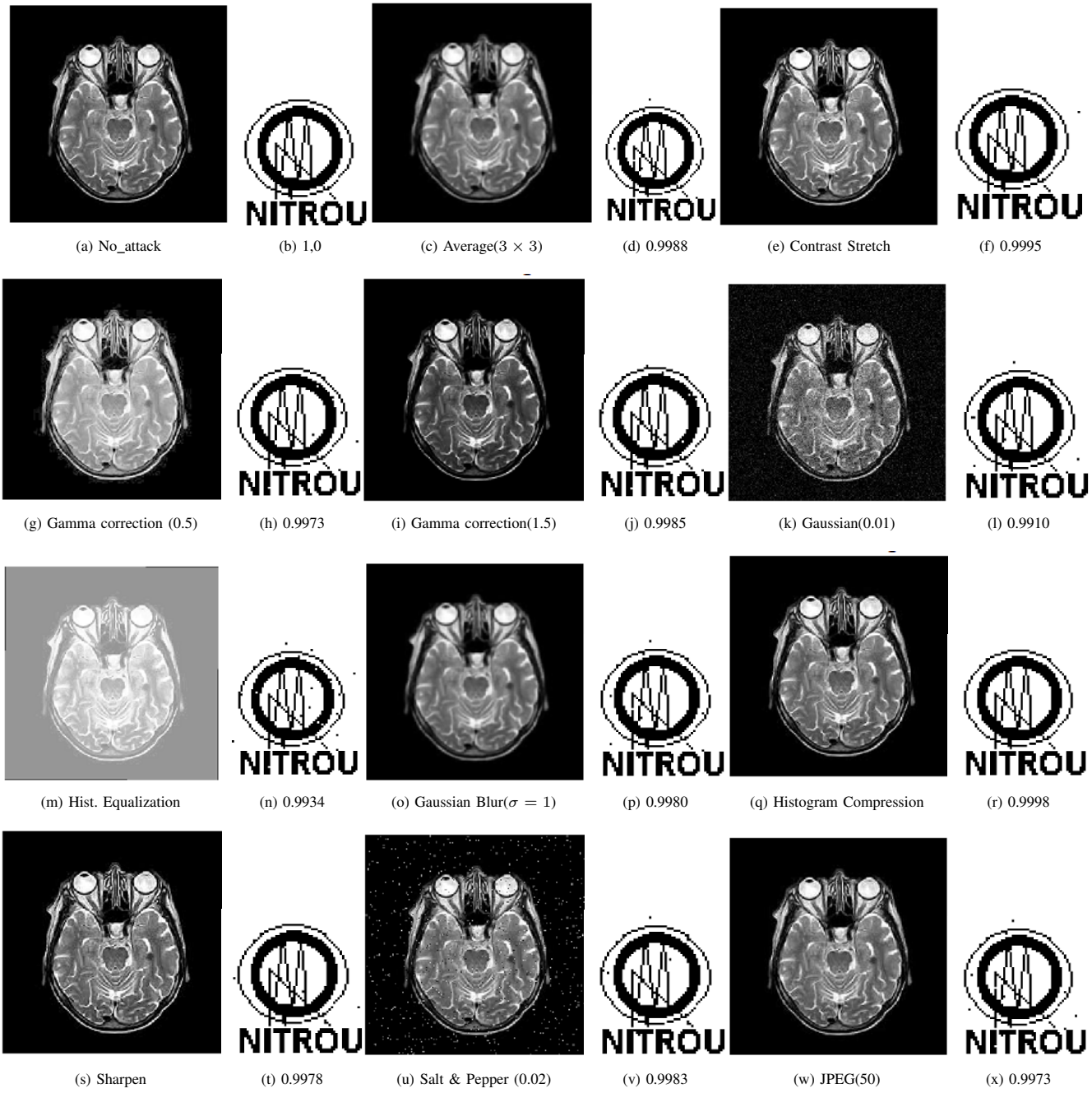
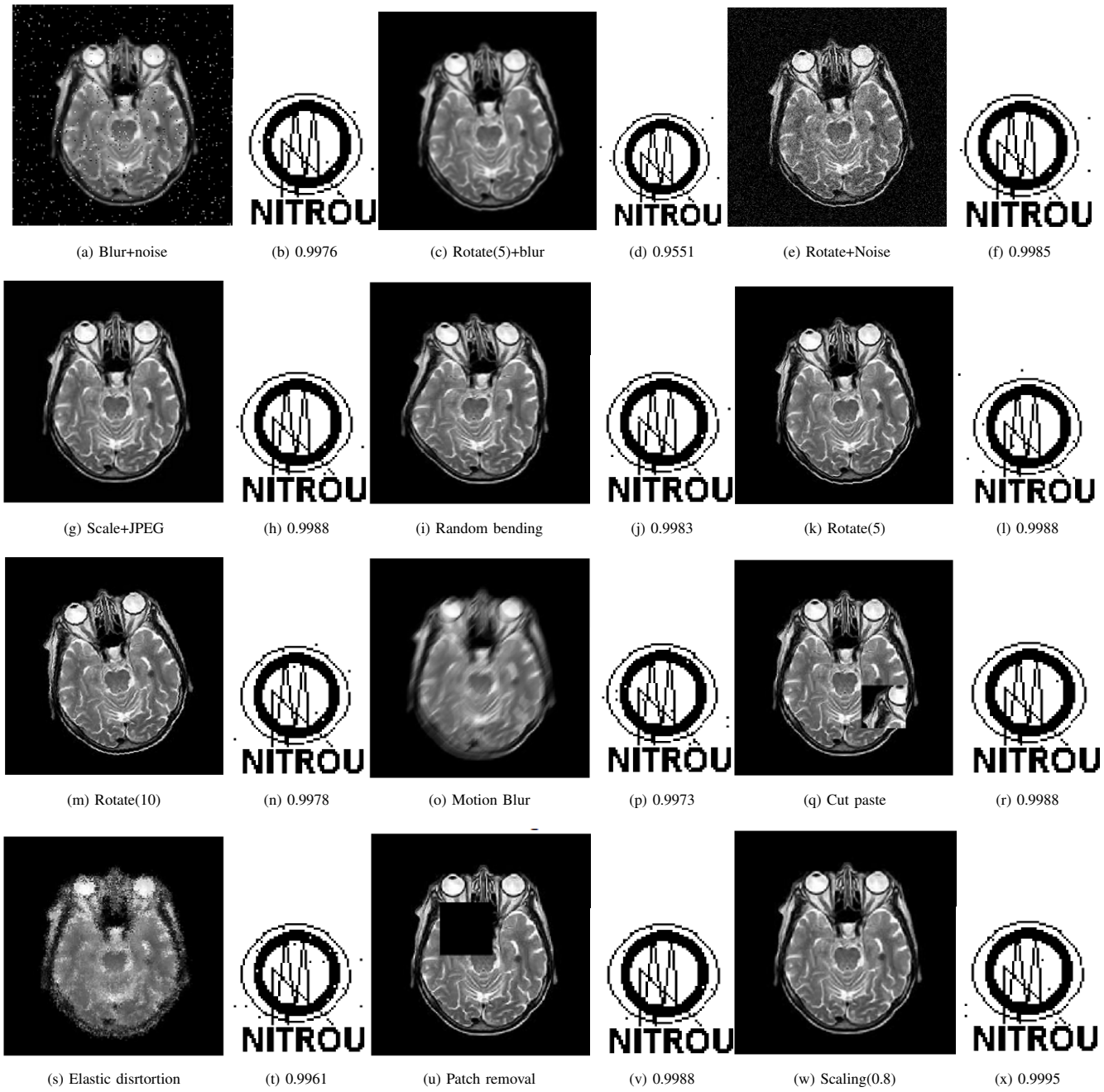


Fig. 5. Signal processing attacks on Brain Image. The SSIM and BER values of the Attacked image and extracted watermarks are calculated. (a) No attack, SSIM=1 (b) BER=0 , (c) Average filter(3×3), SSIM=0.9238 (d) BER=0.0017, (e) Contrast Stretch ,SSIM=0.9987 (f) BER=0.0005, (g) Gamma Correction (0.5), SSIM=0.8837 (h) BER=0.0007, (i) Gamma correction (1.5), SSIM=0.9429 (j) BER=0.0007 (k) Gaussian (0.01), SSIM=0.3045 (l) BER=0.0046, (m) Hist. equalization, SSIM=0.2505 (n) BER=0.0076, (o) Gaussian Blur ($\sigma = 1$), SSIM=0.9182, (p) BER=0.0029, (q) Hist. compression ,SSIM=0.8558 (r) BER=0.0005, (s) Sharpen, SSIM=0.9783, (t) BER=0.0010, (u) Salt & Pepper (0.02), SSIM=0.6746(v) BER=0.0.0020, (w) JPEG (Q=50), SSIM=0.9681, (x) BER=0.0024.

the watermark. This tests the worst-case robustness of the algorithm. Instead of applying one attack at a time, several attacks such as GN0.01, JPEG10, ROT5, and MedF are in a pipeline.

Similarly, a comparison on the imperceptibility test is depicted in Table IV for 10 types of images. It is observed that the PSNR of the proposed method outperforms by a large margin, $\geq$ 10 dB consistently than the other methods.

Fig. 11 shows a spider plot comparison of NCC values under fifteen types of image processing attacks. It is observed that the proposed method maintains NCC values close to unity across all distortions, indicating superior robustness compared with other methods. The scheme in [17] employs a complex feature-extraction process, comprising VGG19, LBP,



(a) Blur+noise (b) 0.9976 (c) Rotate(5)+blur (d) 0.9551 (e) Rotate+Noise (f) 0.9985

(g) Scale+JPEG (h) 0.9988 (i) Random bending (j) 0.9983 (k) Rotate(5) (l) 0.9988

(m) Rotate(10) (n) 0.9978 (o) Motion Blur (p) 0.9973 (q) Cut paste (r) 0.9988

(s) Elastic disrtortion (t) 0.9961 (u) Patch removal (v) 0.9988 (w) Scaling(0.8) (x) 0.9995

Fig. 6. Geometrical combined attacks on Brain image. The SSIM and BER values of the watermarked image and extracted watermarks using proposed scheme (a) Blur+Noise, SSIM=0.6008, (b) BER=0.0024, (c) Rotate (5°)+ Blur, SSIM=0.8962, (d) BER=0.0022, (e) Rotate (5°)+ Noise, SSIM=0.3616, (f) BER=0.0029, (g) Scale + JPEG, SSIM=0.8793 (h) BER=0.0020, (i) Random bending, SSIM=0.6965 (j) BER=0.0017 (k) Rotate (5°, SSIM=0.9728 (l) NCC=0.0010, (m) Rotate (10°), SSIM=0.9723 (n) BER=0.0029, (o) Motion blur, SSIM=0.7674 (p) BER=0.0.0032, (q) Cut and paste, SSIM=0.9539 (r) BER=0.0017, (s) Elastic distortion, SSIM=0.7003 (t) BER=0.0051, (u) Patch removal, SSIM=0.9386 (v) BER=0.0005, (w) Scaling (0.8), SSIM=0.9771, BER=0.0024.

DCT, and DWT. Whereas ZWNet [34] combines stationary wavelet transform (SWT), DarkNet53 deep feature extraction, and Fibonacci Q-matrix encryption. In contrast, our method uses only VGG16+LBP for feature extraction, reducing the complexity of implementation.

To statistically validate the performance improvement, the Wilcoxon signed-rank test was conducted between the proposed method and existing methods using NCC values obtained under fifteen attacks. The obtained p-values are less than 0.05, confirming that the proposed method achieves statistically significant improvements. A comparison of p-values is shown in Table V

To further validate the performance difference among multiple watermarking algorithms, the Friedman test was con-

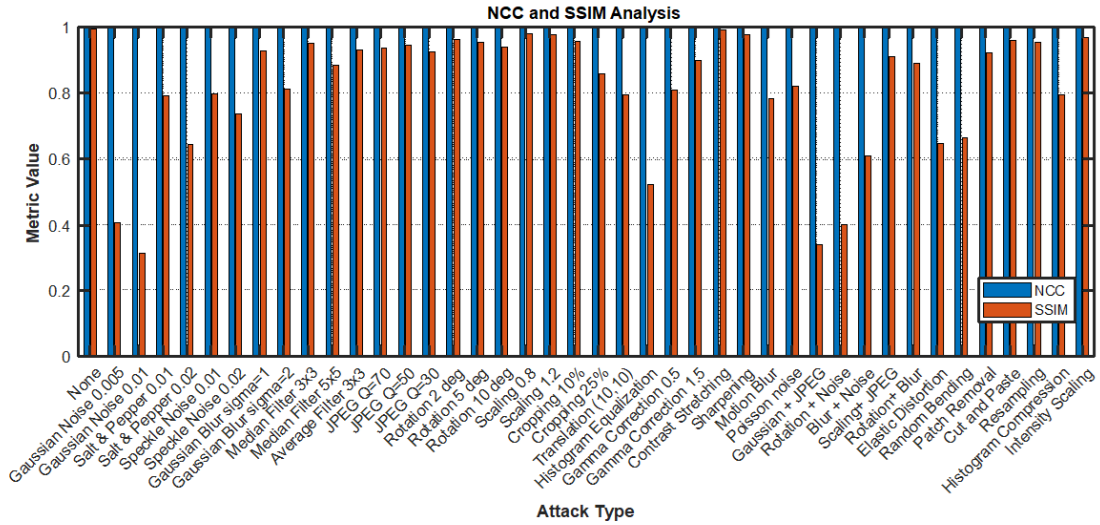


Fig. 7. NCC and SSIM performance under various attacks.

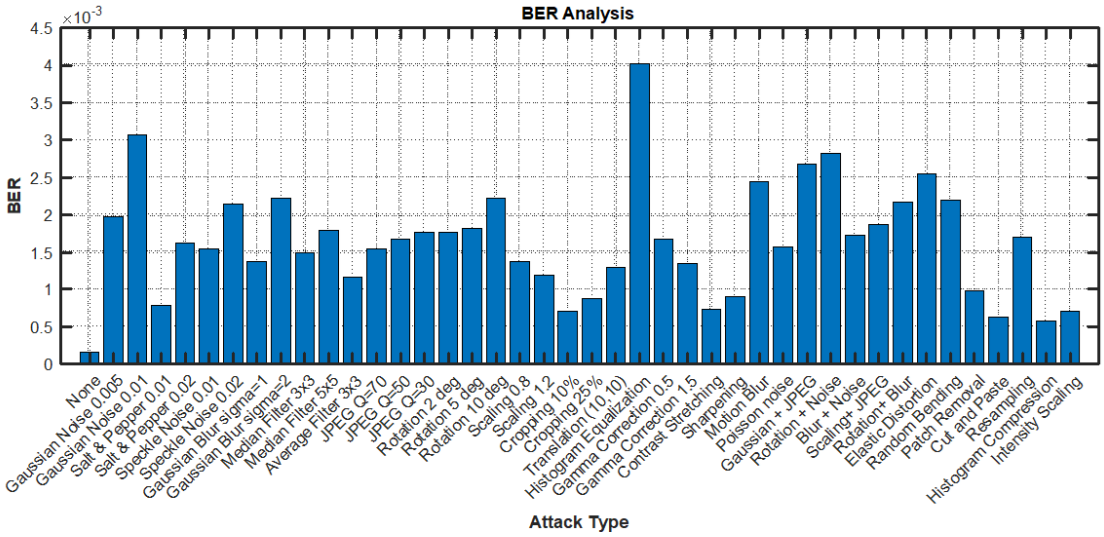


Fig. 8. BER plot under various attacks.

TABLE III. ROBUSTNESS COMPARISON USING NCC UNDER 15 DIFFERENT ATTACKS

Attack	Proposed	[17]	[32]	[33]	ZWNet [34]
GN0.01	0.998	0.995	0.90	0.83	0.992
GN0.05	0.997	0.994	0.89	0.82	0.989
SP0.01	0.997	0.994	0.90	0.83	0.991
SP0.03	0.996	0.993	0.88	0.82	0.988
AvgF	0.998	0.995	0.70	0.96	0.994
MedF	0.998	0.995	0.68	0.95	0.993
GaussF	0.998	0.995	0.69	0.95	0.993
JPEG10	0.996	0.995	0.90	0.92	0.989
JPEG30	0.996	0.995	0.89	0.92	0.990
JPEG50	0.995	0.994	0.88	0.91	0.988
Rot3	0.998	0.996	0.83	0.67	0.994
Rot5	0.998	0.995	0.82	0.66	0.992
Scale0.5	0.997	0.994	0.90	0.92	0.991
Scale1.5	0.997	0.994	0.89	0.91	0.990
Combined	0.998	0.993	0.80	0.90	0.987

TABLE IV. PSNR COMPARISON (DB) BETWEEN THE PROPOSED METHOD AND EXISTING METHODS

Medical Image	Proposed	[17]	[32]	[33]	ZWNet [34]
Brain MRI	58.72	46.31	41.85	39.74	55.92
Chest CT	57.94	45.88	40.92	38.63	55.37
Lung CT	58.21	46.05	41.13	39.22	55.64
Abdominal CT	57.63	45.41	40.55	38.91	55.11
Ultrasound	58.09	45.97	41.02	39.05	55.48

ducted using NCC values obtained under fifteen attacks. The p-value ($\leq$ 0.05) in Table VI indicates that the proposed framework significantly outperforms the compared methods.

Higher NCC under various attacks does not necessarily indicate overfitting or limited variability. The robustness is primarily attributed to 1) ROI selections, 2) APBT low coefficient calculations, 3) LBP feature extraction, and 4) ORB-based geometric alignment. The experiments are conducted across multiple image modalities, introducing variety into the experimental process under 42 different kinds of attacks. In addition, statistical evaluation using the Wilcoxon signed-rank

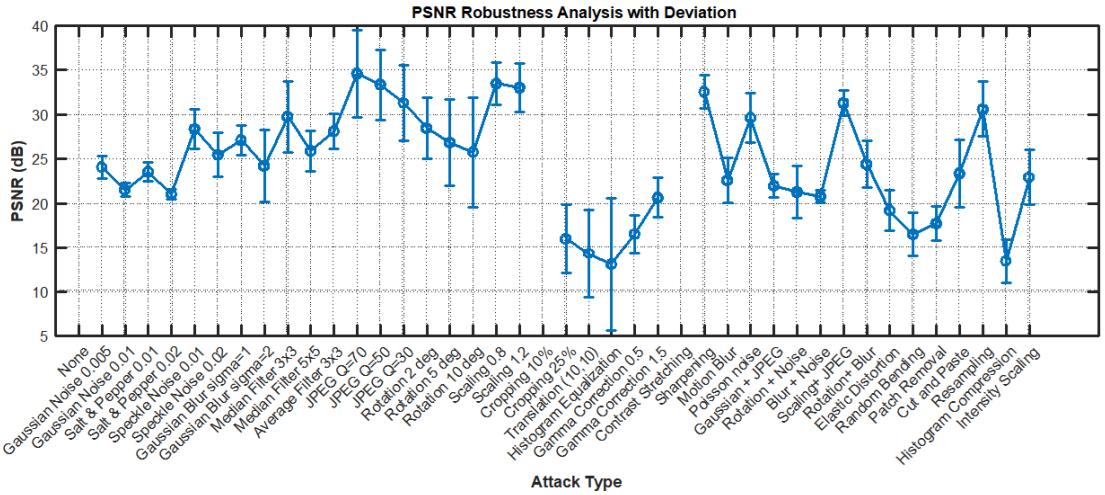


Fig. 9. PSNR performance under various attacks.

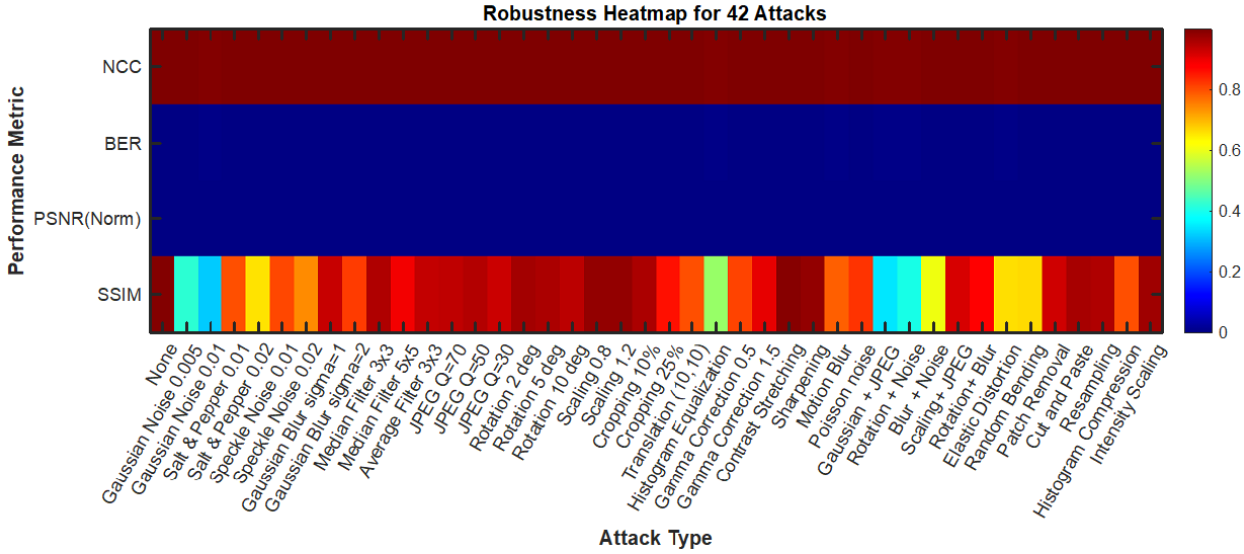


Fig. 10. Heat map plot under various attacks.

TABLE V. WILCOXON SIGNED-RANK AND FRIEDMAN RANK TEST COMPARISON

Comparison	p-value	Average Rank	Significant
Proposed vs [17]	1.2×10^{-3}	1.27	Yes
Proposed vs [32]	3.5×10^{-5}	1.13	Yes
Proposed vs [33]	2.8×10^{-5}	1.07	Yes
Proposed vs ZWNet [34]	8.6×10^{-4}	1.34	Yes

TABLE VI. FRIEDMAN RANKING COMPARISON BASED ON NCC VALUES UNDER 15 ATTACKS

Method	Average Rank
Proposed Method	1.00
Method in [17]	2.87
Method in [32]	3.74
Method in [33]	4.38
ZWNet [34]	2.01

test and the Friedman ranking test confirms that the observed performance improvement is not a consequence of overfitting.

C. Ablation Studies

Table VII illustrates the contributions of individual components of the proposed method towards the average NCC and average BER parameter calculation. The average NCC and BER are evaluated across 10 different image types using the same binary watermark. While the APBT contributes an average NCC of 0.9124, adding LBP yields a gain of 0.0363. The deep-level features of VGG16 only contribute an average NCC of 0.9315. Combining all the features in a single framework contributes a significant average NCC and average BER of approximately 1 and 0, respectively.

D. Computational Complexity and Execution Time

The computational overhead occurs primarily from the pretrained VGG-16 feature extraction, followed by the DNA

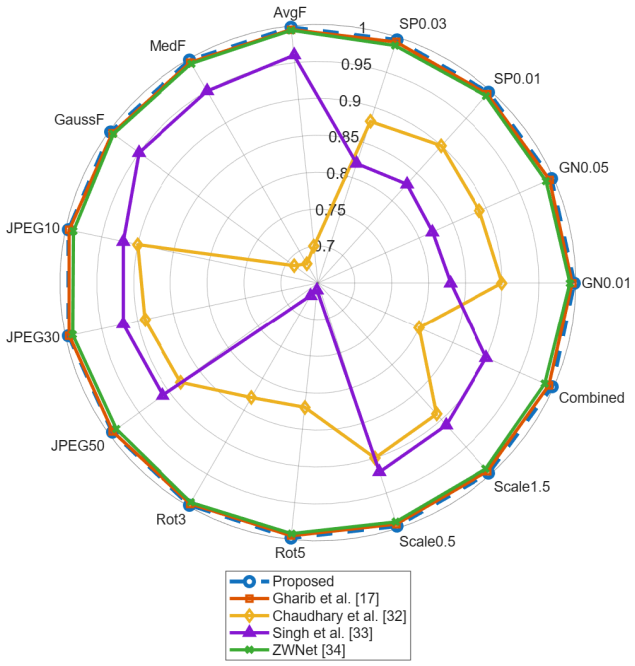


Fig. 11. Spider plot comparison of NCC values under 15 image processing attacks.

TABLE VII. ABLATION STUDY

Features Used	Avg. NCC	Avg. BER
APBT only	0.9124	0.0841
APBT + LBP	0.9487	0.0416
VGG16 only	0.9315	0.0563
APBT + LBP + VGG16	0.9982	0.0006

TABLE VIII. COMPUTATIONAL COMPLEXITY AND MEMORY USAGE COMPARISON

Method	Feature Extraction	Avg Time (s)	Memory (MB)	Complexity
Proposed	APBT+LBP+VGG16	1.24	118	Moderate
[17]	DWT+DCT+VGG19	3.25	214	High
[32]	DWT+HMD+SVD	2.51	176	High

encryption process. However, the proposed scheme is computationally more efficient than most of the existing deep-learning-based watermarking approaches because 1) only stable ROI blocks are selected rather than the whole image blocks, 2) dimensionality reductions are achieved due to low-frequency APBT coefficients, and 3) a pretrained VGG-16 is employed instead of full training. Finally, DNA encryption is applied to compact binary feature vectors rather than full-resolution images. This reduces encryption overhead.

Table VIII shows the average execution time and memory usage of the proposed method in comparison with the other methods. It is observed that the average execution time is nearly 33% and 50% less compared to methods in [17] and [32], respectively, while memory usage is proportionally less. The memory usage is calculated using the MATLAB profiling function by measuring the difference between before and after execution. The VGG16 is not the most lightweight architecture. VGG16 is chosen in our work for its strong feature representation capabilities, architectural simplicity, and widespread adoption. It is used only as a pre-trained feature extractor without end-to-end training. The fused ROI features are processed rather than the entire image, significantly reducing computation. The term 'lightweight' refers to a reduced feature-processing pipeline, achieved through compact fusion of APBT and LBP features rather than using ultra-lightweight architectures such as MobileNet, ShuffleNet, or EfficientNet. Compared with VGG19, ResNet, and Transformer architectures, VGG16 provides a favorable trade-off between memory usage and feature discrimination.

VII. CONCLUSIONS

This paper presented a robust medical image zero-watermarking framework that integrates transform-domain analysis with deep feature extraction. The proposed method employs the All Phase Biorthogonal Transform to capture stable frequency-domain characteristics and utilizes deep features obtained from VGG16 to enhance feature discrimination. Furthermore, ORB-based alignment improves robustness against geometric distortions. Experimental results show that the proposed framework achieves high robustness under various signal processing and geometric attacks. Additionally, the zero-watermarking strategy preserves the diagnostic quality of medical images by avoiding direct modification of the host medical image. Comparative analysis with existing methods and statistical validation using the Wilcoxon signed-rank statistical test and the Friedman ranking test confirm the effectiveness of the proposed approach. Future work will focus on further reducing computational complexity through lightweight deep learning architectures for higher-dimensional medical images.

REFERENCES

[1] M. Yang, J. Li, U. Bhatti, C. Shao, and C. Yen-Wei, "Robust watermarking algorithm for medical images based on non-subsampled shearlet transform and schur decomposition," *Computers, Materials, & Continua*, vol. 75, no. 3, p. 5539, 2023.

[2] R. K. Senapati, B. Biswal, S. Kautish, G. Swain, A. Altameem, A. S. Almazyad, and A. W. Mohamed, "Optimized video watermarking with entropy-aware block selection and modified hénon-map encryption," *IEEE Access*, 2024.

[3] A. M. Wagdarikar and R. K. Senapati, "Design and development of a multiobjective cost function for robust video watermarking using wavelet transform," *Journal of Intelligent Systems*, vol. 28, no. 5, pp. 873–891, 2019.

[4] B. Ahmaderaghi, J. M. Del Rincon, F. Kurugollu, and A. Bouridane, "Perceptual watermarking for discrete shearlet transform," in *European Workshop on Visual Information Processing*. Institute of Electrical and Electronics Engineers Inc., 2014, pp. 1–6.

[5] A. Hadid, "The local binary pattern approach and its applications to face analysis," in *2008 First Workshops on Image Processing Theory, Tools and Applications*, 2008, pp. 1–9.

[6] R. K. Senapati, S. Srivastava, and P. Mankar, "RST invariant blind image watermarking schemes based on discrete Tchebichef transform and singular value decomposition," *Arabian Journal for Science and Engineering*, vol. 45, pp. 3331–3353, 2020.

[7] H. Gao and Q. Chen, "A robust and secure image watermarking scheme using surf and improved artificial bee colony algorithm in dwt domain," *Optik*, vol. 242, p. 166954, 2021.

[8] A. Soualmi, A. Alti, and L. Laouamer, "An imperceptible watermarking scheme for medical image tamper detection," *International Journal of Information Security and Privacy (IJISP)*, vol. 16, no. 1, pp. 1–18, 2022.

[9] M. Roy, D. M. Thounaojam, and S. Pal, "A perceptual hash based blind-watermarking scheme for image authentication," *Expert systems with applications*, vol. 227, p. 120237, 2023.

[10] H. Chaudhary and V. Vishwakarma, "Digital image watermarking recent trends and techniques: A survey," *J. Inf. Optim. Sci*, vol. 45, no. 4, pp. 1051–1059, 2024.

[11] R. E. Arevalo-Ancona and M. Cedillo-Hernandez, "Zero-watermarking for medical images based on regions of interest detection using k-means clustering and discrete fourier transform," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[12] F. Li and Z.-X. Wang, "A zero-watermarking algorithm based on scale-invariant feature reconstruction transform," *Applied Sciences*, vol. 14, no. 11, p. 4756, 2024.

[13] M. Yamni, H. Karmouni, M. Sayyouri, and H. Qjidaa, "Color stereo image zero-watermarking using quaternion radial tchebichef moments," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2020, pp. 1–7.

[14] K. M. Hosny and M. M. Darwish, "Robust color image watermarking using invariant quaternion legendre-fourier moments," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 24 727–24 750, 2018.

[15] M. Magdy, N. I. Ghali, S. Ghoniemy, and K. M. Hosny, "Multiple zero-watermarking of medical images for internet of medical things," *IEEE Access*, vol. 10, pp. 38 821–38 831, 2022.

[16] Z. Xia, X. Wang, C. Wang, C. Wang, B. Ma, Q. Li, M. Wang, and T. Zhao, "A robust zero-watermarking algorithm for lossless copyright protection of medical images: Z. xia et al." *Applied Intelligence*, vol. 52, no. 1, pp. 607–621, 2022.

[17] H. A. Gharib, N. M. Abdelnapi, and K. M. Hosny, "Robust zero-watermarking for color images using hybrid deep learning models and encryption," *Scientific Reports*, vol. 15, no. 1, p. 28906, 2025.

[18] A. Anand, J. Bedi, and I. Rida, "Miwet: Medical image watermarking using encryption and fusion technique," *Computers and Electrical Engineering*, vol. 115, p. 109114, 2024.

[19] Y. Fan, J. Li, U. A. Bhatti, C. Shao, C. Gong, J. Cheng, and Y. Chen, "A multi-watermarking algorithm for medical images using inception v3 and dct." *Computers, Materials & Continua*, vol. 74, no. 1, 2023.

[20] F. Dong, J. Li, U. A. Bhatti, J. Liu, Y.-W. Chen, and D. Li, "Robust zero watermarking algorithm for medical images based on improved nasnet-mobile and dct," *Electronics*, vol. 12, no. 16, p. 3444, 2023.

[21] W. Zhang, J. Li, U. A. Bhatti, J. Liu, J. Zheng, and Y.-W. Chen, "Robust multi-watermarking algorithm for medical images based on googlenet and henon map." *Computers, Materials & Continua*, vol. 75, no. 1, 2023.

[22] M. Sheng, J. Li, U. Bhatti, J. Liu, M. Huang, and C. Yen-Wei, "Zero watermarking algorithm for medical image based on resnet50-dct," *Computers, Materials, & Continua*, vol. 75, no. 1, p. 293, 2023.

[23] B. Han, H. Wang, D. Qiao, J. Xu, and T. Yan, "Application of zero-watermarking scheme based on swin transformer for securing the metaverse healthcare data," *IEEE journal of biomedical and health informatics*, vol. 28, no. 11, pp. 6360–6369, 2023.

[24] B. Han, R. H. Jhaveri, H. Wang, D. Qiao, and J. Du, "Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data," *IEEE journal of biomedical and health informatics*, vol. 27, no. 2, pp. 804–813, 2021.

[25] X. Wang, M. Wen, X. Tan, H. Zhang, J. Hu, and H. Qin, "A novel zero-watermarking algorithm based on robust statistical features for natural images," *The Visual Computer*, vol. 38, no. 9, pp. 3175–3188, 2022.

[26] Z.-X. Hou, C.-Y. Wang, and A.-P. Yang, "All phase biorthogonal transform and its application in jpeg-like image compression," *Signal Processing: Image Communication*, vol. 24, no. 10, pp. 791–802, 2009.

[27] L. Chu, Y. Su, X. Yao, P. Xu, and W. Liu, "A review of dna cryptography," *Intelligent Computing*, vol. 4, p. 0106, 2025.

[28] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer vision using local binary patterns*. Springer Science & Business Media, 2011, vol. 40.

[29] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 Radiography Database," https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, 2020, accessed: May 22, 2026.

[30] M. Nickparvar, "Brain Tumor MRI Dataset," https://www.kaggle.com/dsv/2645886, 2021, accessed: May 22, 2026.

[31] S. Kukreja, G. Kasana, and S. S. Kasana, "Extended visual cryptography-based copyright protection scheme for multiple images and owners using LBP–SURF descriptors," *The Visual Computer*, vol. 37, no. 6, pp. 1481–1498, 2021.

[32] H. Chaudhary, P. Garg, and V. P. Vishwakarma, "Enhanced medical image watermarking using hybrid dwt-hmd-svd and arnold scrambling," *Scientific Reports*, vol. 15, no. 1, p. 9710, 2025.

[33] K. N. Singh, O. P. Singh, A. K. Singh, and A. K. Agrawal, "Watmif: Multimodal medical image fusion-based watermarking for telehealth applications," *Cognitive Computation*, vol. 16, no. 4, pp. 1947–1963, 2024.

[34] M. M. Abdel-Aziz, N. A. Lashin, H. M. Hamza, and K. M. Hosny, "Zwnet: Darknet53-based zero watermarking method for authentication of medical images inspired by fibonacci q-matrix and stationary wavelet transform," *Biomedical Signal Processing and Control*, vol. 100, p. 107044, 2025.

A Privacy-Aware Federated Hybrid Model for Multimodal Mental Health Analysis

Yusra, Riaz UlAmin
Smart Systems Research Lab, University of Okara, Pakistan

***Abstract*—As mental health disorders such as stress, anxiety, depression, and post-traumatic stress disorder (PTSD) affect a substantial part of the world population, current diagnostic methodologies are still centralized, subjective, and sensitive to privacy concerns. To mitigate these limitations, this study presents a new framework for multimodal mental health classification within a privacy-preserving federated learning framework learning using electroencephalography (EEG), electrocardiography (ECG) and galvanic skin response (GSR) signals. Furthermore, we propose a hybrid deep learning architecture, which combines CNN-LSTM-Transformer blocks to effectively learn spatial, temporal, and long-range dependencies within physiological signals.After preprocessing the cleaned data through artifact removal, band-pass filtering, normalization and multimodal feature fusion signal quality is improved. The proposed model is trained in a federated setting with multiple clients for decentralized training without sharing raw data allowing it to preserve privacy and communication efficiency supporting non-IID data extensions. We evaluate on two datasets, SAM40 (stress detection) and DAPS (anxiety, depression, and PTSD classification). The proposed framework achieved 97% accuracy on SAM40 and more than 96% accuracy on DAPS. Comparative assessments with recent federated and centralized methods validate its strength in multimodal fusion and robust feature exploitation. These results demonstrate the possibility of a general framework for designing privacy-preserving and efficient mental health monitoring systems that can support both Clinical and Wearable-device applications.**

***Keywords*—Mental health detection; hybrid CNN-LSTM-transformer; federated learning; multimodal physiological signals**

I. Introduction

Mental health illnesses, like depression, stress, and anxiety, affect one billion people each year and create significant social, economic, and healthcare burdens. Depression, stress, and anxiety disorders are some of the top causes of disability, and loss of healthy and productive years of life [1]. Depression is a major cause of suicide and loss of work productivity and costs the world economy 1 trillion dollars a year [2].

Mental disorders must be identified and assessed to provide intervention. Current methods, such as interviews interviews and self-reported questionnaires are subjective, time-consuming, and impractical. Measuring physiological states can be used to indicate mental health because they are objective. Physiological signals continuously monitored and measured. EEG records brain activity, ECG measures heart activity due to stress or anxiety, and GSR measures the arousal from the autonomic nervous system [3]. While machine learning techniques can efficiently analyze given problems, centralized methods require collecting personal and sensitive data, raising serious privacy concerns. Another important point is that most studies focus on analyzing a single type of signal, and ignores other potentially useful signals. Real-world data that can be collected from hospitals, clinics, and wearable devices are also non-independent and identically distributed (or non-IID), which makes it difficult for models to generalize effectively [4].

To mitigate those problems, advanced preprocessing and feature extraction is done before federated training. EEG signals are filtered to a desired band, artifacts are removed, and other signals are divided into segments. For processed ECG signals, we determine R-peaks and extract important features such as heart rate and heart rate variability (HRV) from our filtered signals. GSR signals are smoothed and normalized, and we identify arousal peaks. These methods minimize added noise in data, make the characteristics and inputs we are analyzing uniform, and improve feature extraction through cross-feature analysis in time and space across different signals. [5].

Federated Learning (FL) allows for training models in a distributed manner across devices in a privacy-preserving manner, which allows which allows for better compliance with privacy regulations. Using multiple modalities in a federated system improves detection performance by enabling the capture of spatial, temporal, and cross-modal dependencies. We present a novel CNN-LSTM-Transformer architecture where the convolutional layers capture the spatial patterns, the LSTM layers capture the temporal patterns, and the Transformer layers capture long-range and cross-modal dependencies. The framework has capacity to further integrate privacy-preserving techniques such as Differential Privacy and Secure Aggregation to help keep distributed training secure and ensure that model updates are safely combined in the federated learning setting [6].

As illustrated in Fig. 1, the AI system supports anxiety disorder management using multiple modalities.

The main contributions of this work are:

- We explore a multimodal federated learning framework integrating EEG, ECG, and GSR for enhanced and robust mental health detection.
- We developed a CNN-LSTM-Transformer model to illustrate hierarchical feature extraction from preprocessed physiological signals.
- We remove noise and artifacts from EEG, ECG, and GSR signals to improve signal integrity and feature reliability.
- We comprehensively evaluate the proposed approach in terms of accuracy, F1-score, convergence rate, communication efficiency, and privacy preservation.

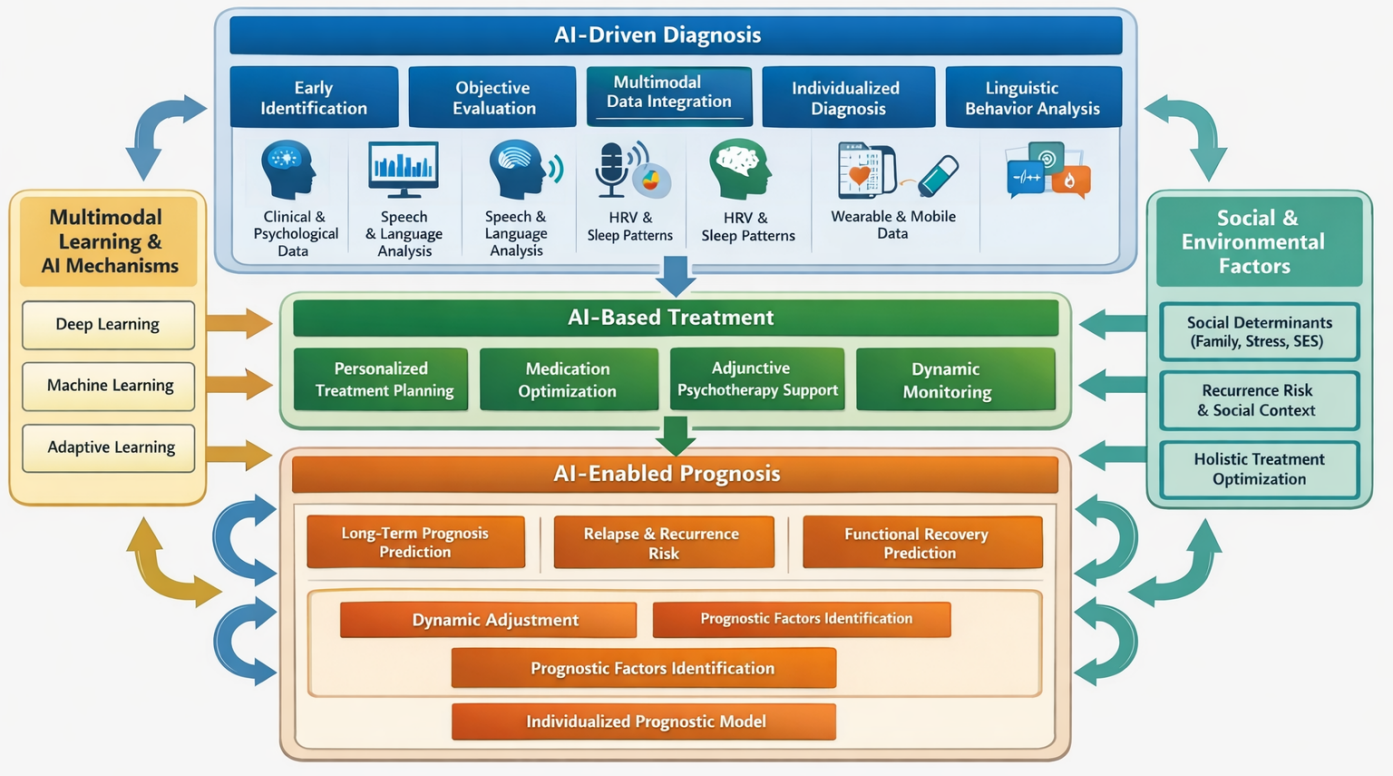


Fig. 1. AI-Driven conceptual framework for anxiety disorder integrating diagnosis, treatment, and prognosis with multimodal data, machine learning, and social-environmental factors.

II. Literature Review

Recently, federated learning (FL) has emerged to enhance privacy-preserving mental health and emotion recognition systems by enabling decentralized training without sharing raw physiological data. Although there have been these advances, the majority of the current studies still have drawbacks in terms of multimodal fusion, robustness against non-IID data distribution and scalability for real clinical applications. In their proposed privacy-preserving federated framework for depression relapse prediction from EEG signals and clinical symptom scores, Yasin et al. [7] introduced a method that leverages machine learning to predict depression relapse by analyzing electroencephalogram (EEG) signals and symptom scores. The model uses explainable AI methods like Layer-wise Relevance Propagation (LRP) and Shapley values, and it has an accuracy of 92%. However, the method does not consider multimodal physiological integration, and is only applied to single-modality EEG data.

Dubey et al. [8] proposed a multimodal federated learning system with CNN and LSTM networks under various privacy constraints. The method remains sensitive to data heterogeneity and is not very adaptive to various client distributions as it achieves only around 92% accuracy. Simic et al. [9] presented a federated multimodal emotion recognition system based on CNN models for audio and visual feature extraction. The model uses decision-level fusion, which restricts the ability of deep cross-modal feature interaction resulting in limited affective state modeling performance, although FedAvg can enhance the generalization of the model.

Gupta et al. [10] proposed an EEG-audio-based federated multimodal system for the early detection of the Major Depressive Disorder using Bi-LSTM based fusion. Although the model performed well with 91.9% accuracy, the extent to which the model has been validated on real-world heterogeneous datasets is limited, which results in a lack of clinical generalization.

Almadhor et al. [11] studied stress detection in the context of wearable electrodermal activity (EDA) signals in a federated learning framework. The system provides 86.82% accuracy but under device variations in the devices and limited diversity in the datasets, the accuracy of the system drops.

Kumar et al. [12] proposed a federated transfer learning approach (FTL-Emo) for EEG-based emotion recognition. The model obtains an F1-score of 92-94%, but does not generalize well across datasets due to distribution shifts between the training and testing domain. Federated reinforcement learning combined with machine unlearning was used in healthcare monitoring and anomaly detection by Shaik et al. [13]. The framework offers flexibility and privacy, but it has high computational complexity, which restricts its use in real-time applications.

Farooq et al. [14] proposed a federated learning approach involving local classifiers with a meta-classifier for the detection of autism spectrum disorders. When applied to children, the method achieved 98% accuracy, but decreased to 81% for adults, thereby showing poor generalisation across age groups.

More recently, Pradeep et al. [15] investigated cross-modal

TABLE I. COMPREHENSIVE LITERATURE REVIEW OF FEDERATED LEARNING AND HYBRID AI APPROACHES IN MENTAL HEALTH AND EMOTION RECOGNITION

Reference	Objective	Methodology	Input Data	Hybrid	AI/ML	Limitations	Accuracy (%)	Bench-marked
[7]	Early prediction of depression relapse with privacy preservation	Federated learning with EEG features and clinical scores; explainable AI	EEG + symptom scores	✓	✓	Computational demand, EEG quality dependent	92	✓
[8]	Privacy-preserving mental health prediction	Federated learning with CNN + LSTM, multimodal data integration, privacy	EEG, ECG, GSR, behavioral signals	✓	✓	Data heterogeneity, privacy trade-offs	92	✓
[9]	Privacy-preserving multimodal emotion recognition on edge devices	CNN-based audio + video model with federated learning (FedAvg)	Facial video frames + Audio spectrograms	✓	✓	Sensor variability, model interpretability	92.86–94.05	✓
[10]	Early MDD detection with privacy	Federated multimodal deep learning (EEG + audio, Bi-LSTM)	EEG + audio signals	✓	✓	Limited real-world validation	91.9	✓
[11]	Privacy-preserving stress detection	Federated DNN (with SMOTE and normalization)	Wearable electrodermal activity (WESAD dataset)	✓	✓	Limited sample size, device variability	86.82	✓
[12]	Privacy-preserved automatic emotion recognition	Federated transfer learning (CNN + FL)	EEG biomarkers (DEAP + K-EmoCon)	✓	✓	Limited cross-dataset generalization	92–94	✓
[13]	Healthcare monitoring using FRL and machine unlearning	Federated reinforcement learning (FedStack, FRAMU)	Multimodal sensor	✓	✓	Computational complexity, data heterogeneity	-	✓
[14]	Early ASD detection with privacy	Federated learning with local LR + SVM and meta-classifier	Four ASD feature datasets	✓	✓	Limited adult dataset generalization	98 (children), 81 (adults)	✓
[15]	Cross-modal brain–heart interaction modeling using EEG and ECG feature mapping	EEG–ECG cross-modal feature learning with regression-based mapping	EEG + ECG signals	✓	✓	No federated learning, no privacy-preserving framework, limited scalability	93.4	✗
Proposed Method	Hybrid CNN-LSTM-Transformer with noise removal for anxiety detection	CNN, LSTM, Transformer, artifact removal, noise filtering	EEG ,ECG,GSR signals	✓	✓	Computationally intensive, EEG variability	97.37	✓

brain–heart interaction modeling using EEG and ECG signals. The study was found to be effective for physiological feature mapping across modalities, but lacks of federated learning or privacy-preserving distributed training, making it less applicable in real-world decentralized healthcare settings.

The literature summarized in Table I shows that recent studies have explored the use of federated learning techniques for mental health and emotion recognition. Although current approaches help privacy of data, face challenges such as limited generalization, weak multimodal fusion, data sensitivity, and lack of clinical validation in real-world applications.

The proposed method, however, combines the three signals — EEG, ECG and GSR through a hybrid CNN–LSTM–Transformer network with improved feature fusion and noise reduction. This enables the model to capture spatial, temporal and long-range dependencies and maintain robustness in different federated environments, that increases the performance and scalability of real-world mental health monitoring systems.

III. RESEARCH METHODOLOGY

This research develops a novel, multi-modal, federated learning based, mental health detection framework that prioritizes and balances privacy and scalability. The system fuses EEG, ECG, and GSR data and utilizes Convolutional Neural Networks, Long Short-Term Memory networks and Transformers, allowing to capture the spatio-temporal and attention driven features. This staircase approach facilitates the overall characterization and modeling of stress, anxiety, and depression related multi-dimensional physiological responses.

A. Data Collection and Understanding

In this research, the DAPS and SAM40 datasets, which have been incorporated into clinical and wearable health monitoring systems, involve several datasets sampling a wide range of mental states and multiple-spectral modalities.

The SAM40 dataset contains EEG wave recordings of 40 participants who were volunteers in a convenience sample of a greater research project (14 females, 26 males; average age 21.5 years) who were treated to a series of cognitive task activations designed to produce and control stress states. Tasks included were Stroop activities, mental arithmetic, mirror image processing and formal relaxation. Active data collection utilized a 32-channel Emotiv Epoc Flex EEG headset which adhered to the international 10–20 system spatial configuration, where Fp1, Fp2, F3, F4, C3, C4, O1, O2 and a few other electrodes were recorded. Signals were collected and within the resultant 2.75 minutes of quiet recording, were arranged in 25 second, non-overlapping intervals. Each resultant interval epoch data was represented as matrix [16], [17].

$$X^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ \vdots \\ x_C^{(t)} \end{bmatrix} \in \mathbb{R}^{C \times T}, \quad C = 32,\ T = 25 \times 128 \qquad (1)$$

where, $x_c^{(t)}$ corresponds to the time-series signal from channel c for epoch t, with C as the total number of EEG channels and T as the number of samples per epoch. Our collection has both continuous raw and artifact-corrected EEG signals. This feature makes SAM40 a flexible dataset as to the study of brain activities associated with varying stress levels and the use of machine learning tools.

DAPS dataset (custom or proprietary) comprises EEG and GSR signals of GSR and EEG collected from participants having depression, anxiety, or PTSD, as well as healthy controls. EEG is collected with 16–128 channels deploying a high-density clinical or wearable EEG system, GSR is obtained from finger/palm electrodes, and ECG is captured optionally via chest leads or wearable sensors. Data collection occurs during resting-state (eyes open/closed) and emotion or task

induced conditions where clinical scales (e.g. BDI, HAM-A) and interactivity and valence ratings are used to derive labels. Each of the modalities is defined as follows:

$$\mathbf{X}_{EEG}^{(t)} \in \mathbb{R}^{C_{EEG} \times T}, \quad \mathbf{X}_{GSR}^{(t)} \in \mathbb{R}^{C_{GSR} \times T}, \quad \mathbf{X}_{ECG}^{(t)} \in \mathbb{R}^{C_{ECG} \times T} \tag{2}$$

Data are distributed in a heterogeneous, non-IID manner across clients:

$$D = \bigcup_{k=1}^{K} D_k, \quad D_i \cap D_j = \emptyset \text{ for } i \neq j \tag{3}$$

This equation shows that the overall dataset D is the combination of client datasets D_k and that each client's data is disjoint, meaning no clients get to see the same samples. This arrangement mirrors practical federated learning situations with non-IID data distributions.

This organization of datasets assists in providing a strong basis for multimodal mental health detection and allows for the development of privacy-preserving and scalable models.

B. Signal Preprocessing and Feature Engineering

In order to thoroughly assess various mental health conditions based on several physiological signals, a careful range of steps involving a preprocessing pipeline predicated on a theory and the crafting of various features must be devised and executed. Although EEG, ECG, and GSR signals contain substantial amounts of noise and artifacts, they may contain a lot of useful information, especially with regard to inter-subject variance. When performed properly, feature extraction and the noise and artifacts of a signal, termed preprocessing, should be done to obtain a signal that accurately represents the underlying physiological and cognitive states, and this will enable precise modeling.

EEG records electrical activity as generated by postsynaptic potentials due to the synchronous firing of neurons. There are different cognitive and emotional states reflected in varying ranges of delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) waves. However, it's worth noting that EEG is particularly vulnerable to artifacts due to eye and muscle movement, as well as body motion [18]. This is why band-pass filtering is employed in order to retain meaningful activity within the oscillations. The electrical signals generated by the heart (ECGs) are also filtered in order to remove baseline wander and motion artifacts to more accurately quantify the variability in heart rate (HRV), a well-established marker of stress and anxiety, and of autonomic nervous system activity [19]. GSR signals, which capture changes in skin conductance that are due to nervous system activity, are filtered to extract the stress, emotional reactivity, and arousal-response-associated components. The mathematically conveyed expression associated with the artifact-free signal is:

$$\tilde{x}_c(t) = \mathcal{F}_{\text{filter}}(x_c(t)) \tag{4}$$

where, $\tilde{x}_c(t)$ is the preprocessed signal at channel c, $x_c(t)$ is the raw signal, and $\mathcal{F}_{\text{filter}}$ represents the corresponding artifact removal filter.

Physiological data often have samples missing due to sensor malfunctions or motion artifacts. Interpolation is employed to keep the time series continuous, so as to not introduce bias into the statistical and spectral data analyses.

$$x_c(t) = \frac{x_c(t-1) + x_c(t+1)}{2}, \quad \text{if } x_c(t) \text{ is missing} \tag{5}$$

Standardization and normalization are critical to mitigate inter-subject and inter-device variability [20]. Signals are standardized to zero mean and unit variance:

$$\begin{aligned} \tilde{x}_c(t) &= \frac{x_c(t) - \mu_c}{\sigma_c}, \\ \mu_c &= \frac{1}{T}\sum_{t=1}^{T} x_c(t), \\ \sigma_c &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_c(t) - \mu_c)^2} \end{aligned} \tag{6}$$

Here, μ_c and σ_c are the mean and standard deviation for the signal of each channel c. Lastly, for additional variability mitigation, normalization can rescale signals to a range of [0,1]:

$$x_c^{\text{norm}}(t) = \frac{x_c(t) - x_c^{\min}}{x_c^{\max} - x_c^{\min}} \tag{7}$$

Feature extraction is based on the realization that psychological conditions are expressed as patterned changes on multiple time scales, spectral compositions, and physiological changes of various durations. Characteristics of the time domain, such as mean, variance, standard deviation, as well as HRV, are required to assess the overall dynamics of a signal:

$$\text{Mean: } \mu = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \text{Variance: } \sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)^2 \tag{8}$$

The Fast Fourier Transform (FFT) is utilized to derive frequency-domain characteristics through quantifying power in specific frequency bands related to stress, attention, or relaxation as follows:

$$X(f) = \sum_{t=0}^{T-1} x_t e^{-j2\pi ft/T} \tag{9}$$

Features in the time-frequency domain corresponded to the non-stationary and transient characteristics [21], which are relevant to the modeling of stress and emotions, and are obtained through the use of Discrete Wavelet Transform (DWT) as:

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int x(t)\psi\left(\frac{t-b}{a}\right) dt \tag{10}$$

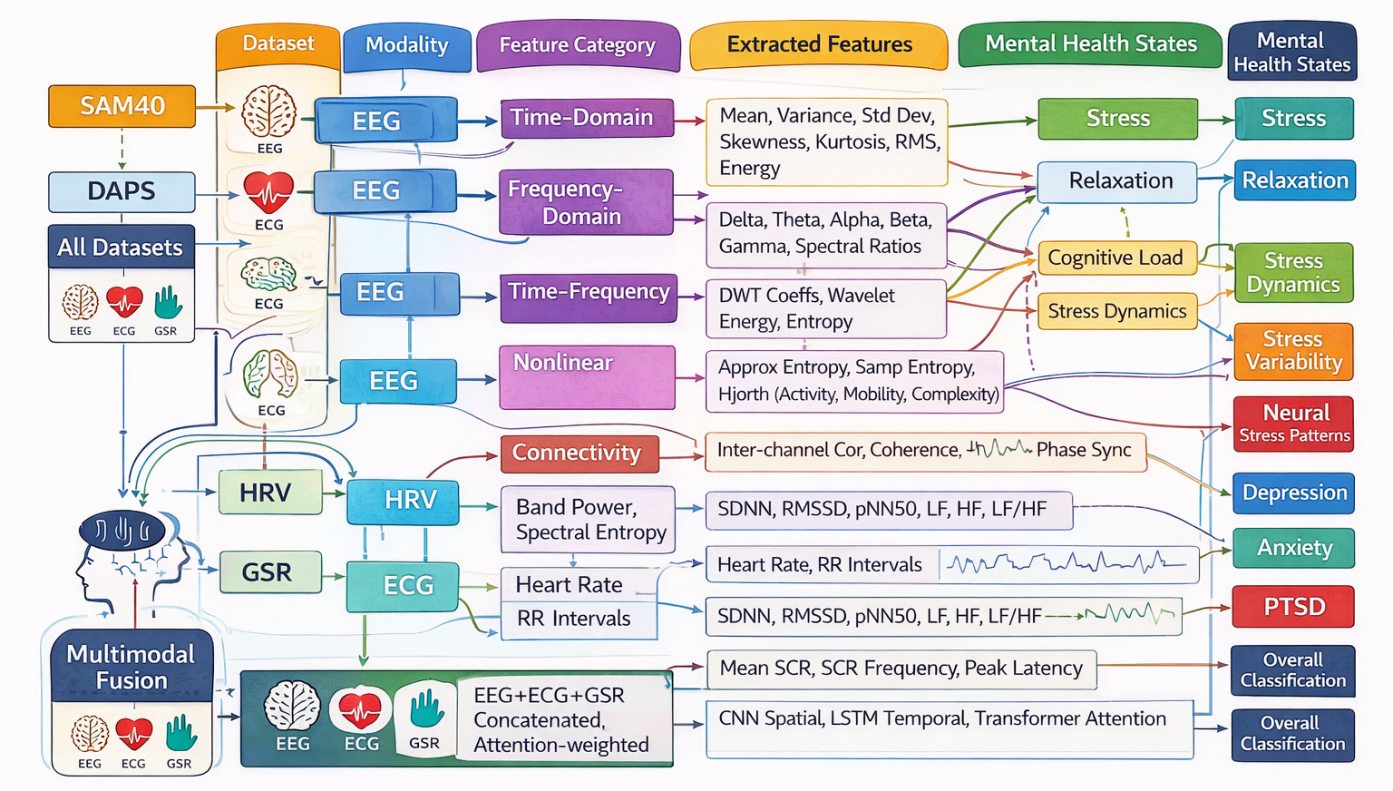


Fig. 2. Correlation of multimodal physiological features with mental health states.

where, a, b are scale and translation respectively and $\psi(t)$ is the mother wavelet.

To take advantage of the complementary information across modalities, the EEG, ECG, and GSR features are concatenated into a single, unified representation:

$$\mathbf{F} = [\mathbf{F}_{EEG}\ \mathbf{F}_{ECG}\ \mathbf{F}_{GSR}] \tag{11}$$

An attention-based weighting mechanism is utilized to emphasize information that is relevant while diminishing information that is redundant or of less relevance:

$$\mathbf{F}_{\text{att}} = \mathbf{F} \odot \alpha, \quad \sum_i \alpha_i = 1 \tag{12}$$

where, α defines learned attention weights and where $\odot$ represents a Hadamard product. This makes sure that the notable physiological patterns contributing to the stress, anxiety, and/ or depression are accentuated in the final feature representation.

There is a feature set providing a rich, complementary, and physiologically meaningful representation due to the combination of artifact-free, uniform, and standardized signals with temporal, spectral, and time-frequency features, which are of distinguished discriminative value, along with the modalities being fused with an attention mechanism. This also strengthens the capabilities of hybrid CNN-LSTM-Transformer models in identifying minute changes in mental health conditions. This makes the technique highly recommended for privacy-preserving federated learning systems as it is as applicable as possible.

Fig. 2 contains the total collection of features of the time domain, frequency domain, time-frequency domain, nonlinear domain, and features of connectivity and of deep learning obtained through the EEG, ECG, and GSR signals from the SAM40 and DAPS datasets. These features are varied in that each of the divisions captures a certain physiological characteristics that correlate to stress, anxiety, depression, and PTSD which aids in effective multimodal mental health classification.

IV. Hybrid CNN-LSTM-Transformer Model Architecture

The Hybrid CNN-LSTM-Transformer model is meant to accurately detect and classify mental health conditions by extracting spatial, temporal, and attention-based features from multimodal physiological signals such as EEG, ECG, and GSR. Specifically, the CNN module is meant to handle the spatial correlations present in the inputs $X \in \mathbb{R}^{T \times F}$, $X \in \mathbb{R}^{T \times F}$, where, T is the number of timesteps and F is the number of features. CNN process begins with one convolutional layer using 64 filters of the 3-sized kernel and generates the following feature maps.

$$H^{(1)} = \text{ReLU}(W^{(1)} * X + b^{(1)}) \tag{13}$$

Subsequently, we perform batch normalization and average pooling to decrease the dimension of the data.

$$H_{\text{pool}}^{(1)} = \max_{i \in \text{pool_size}} H_i^{(1)} \tag{14}$$

A second convolutional layer with 128 filters applies.

$$H^{(2)} = \text{ReLU}(W^{(2)} * H_{\text{pool}}^{(1)} + b^{(2)}) \tag{15}$$

This is then followed by the application of batch normalization and pooling operations; we take the outcome feature maps and flatten them into a single spatial feature vector, denoted as F_{CNN}.

The LSTM module is used to capture temporal dependencies according to the following gating equations at each time step t:

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{16}$$

where, i_t, f_t, o_t are the input, forget, and output gates, c_t is the cell state, and h_t is the hidden state. The first LSTM layer with 128 units returns sequences, followed by a dropout of 0.3, and the second LSTM layer with 64 units outputs the final temporal feature vector $F_{\text{LSTM}} = h_T$.

The Transformer module applies self-attention to capture long-range dependencies:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

where, $Q = HW_Q, K = HW_K, V = HW_V$. Multi-headed attention is concerned with merging different attention components:

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W_O \tag{18}$$

The use of residual connections as well as the application of layer normalization are as follows:

$$H' = \text{LayerNorm}(H + \text{MultiHead}(H)) \tag{19}$$

followed by another feed-forward network:

$$\text{FFN}(H') = W_2 \cdot \text{ReLU}(W_1 H' + b_1) + b_2 \tag{20}$$

and dropout and global average pooling to get the attention-based feature vector:

$$F_{\text{Trans}} = \frac{1}{T}\sum_{t=1}^{T} \text{FFN}(H')_t \tag{21}$$

The outputs of CNN, LSTM and Transformer are concatenated into a hybrid feature vector:

$$F_{\text{Hybrid}} = [F_{\text{CNN}} \parallel F_{\text{LSTM}} \parallel F_{\text{Trans}}] \tag{22}$$

and this is input into fully connected layers consisting of a ReLU activation, batch normalization, and dropout. In particular, a dense layer of 256 units, batch normalization, and a 0.4 dropout, followed by a dense of 128 units, and a final softmax layer gives the prediction:

$$\hat{y} = \text{Softmax}\Big(W_3 \cdot \text{ReLU}(\, W_2 \cdot \text{ReLU}(W_1 F_{\text{Hybrid}} + b_1) + b_2) + b_3\Big) \tag{23}$$

across n_{classes}. The early recognition of stress, anxiety, and PTSD, while maintaining stability and generalization, is made possible because this hybrid architecture seamlessly integrates spatial, temporal, and attenTional characteristics.

V. Federated Learning Framework with Privacy Preservation and Performance Evaluation

In order to achieve safe and cooperative model training, the proposed hybrid CNN-LSTM-Transformer model is integrated into a client-server federated learning environment. Each client k trains the model on its private dataset D_k and generates the following local model updates:

$$\theta_k^{t+1} = \text{Train}(\theta^t, D_k) \tag{24}$$

where, θ^t is the global model parameters at t-th round. Those updates, as opposed to the raw data, are communicated to the central server, which accumulates the updates from the clients using a weighted average according to the FedAvg algorithm:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{t+1} \tag{25}$$

where, n_k is the number of data samples from client k and $n = \sum k = 1^K nk$ is the global sample size held by all clients. To mitigate the non-IID and heterogeneous data distributions problem, adaptive η_k learning rates and unique local model parameters in the form of personalized layers are used to ensure a solid convergence.In order to mitigate the non-IID data heterogeneity across federated clients, FedProx-inspired regularization and personalized local adaptation layers are added to minimize client drift and promote stability of convergence.

A number of privacy preserving mechanisms are assumed to protect sensitive physiological data. Differential Privacy (DP) is realized through DP-SGD, which adds noise $\mathcal{N}(0, \sigma^2)$ to the gradients $\nabla\theta_k$ in order to make it impossible to reconstruct the private data:

$$\tilde{\nabla}\theta_k = \nabla\theta_k + \mathcal{N}(0, \sigma^2) \tag{26}$$

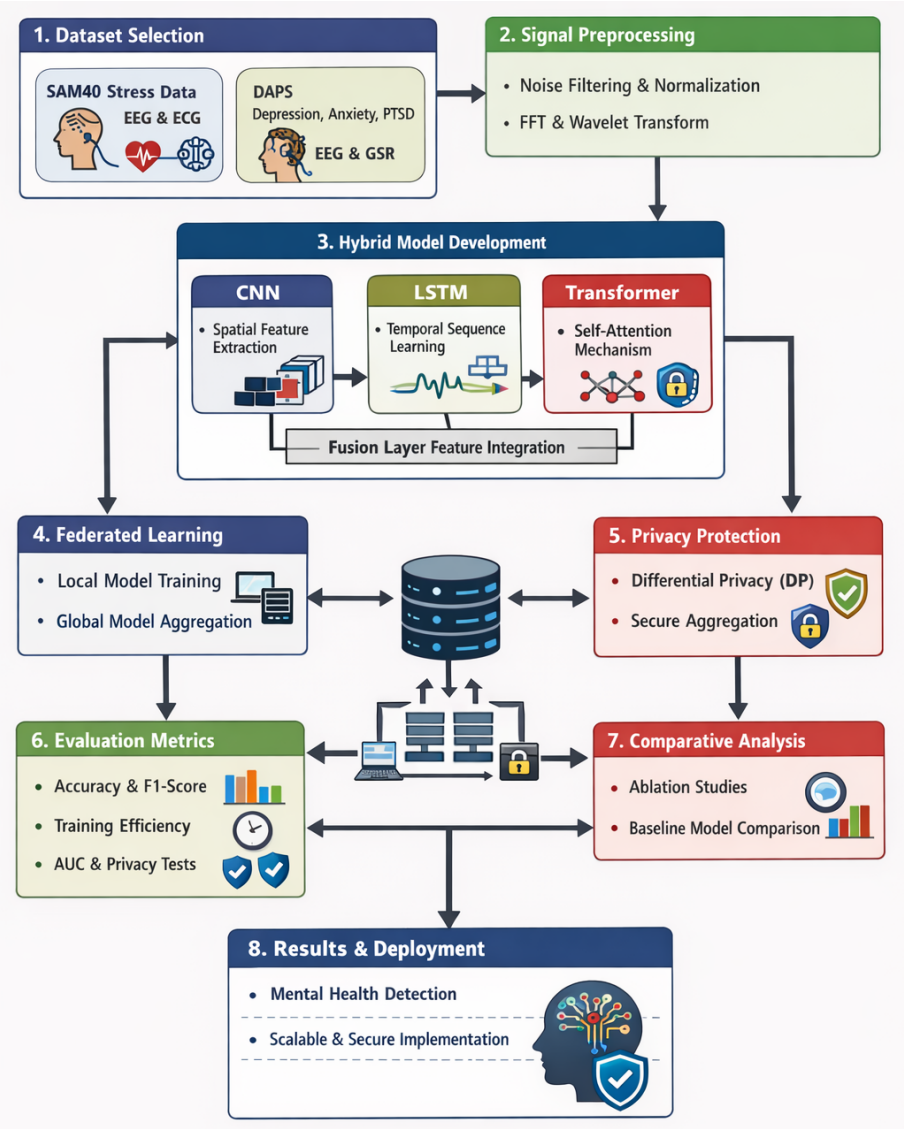


Fig. 3. Proposed federated learning–based hybrid deep learning framework for mental health disorder detection.

Secure aggregation ensures that individual client contributions remain confidential, while adversarial robustness evaluations quantify resilience against attacks such as model inversion, poisoning, and gradient leakage.

The framework is evaluated comprehensively across predictive performance, efficiency, and privacy [22]. Predictive performance metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Training and communication efficiency are analyzed via local training time t_k, global convergence rate, and communication overhead:

$$C = \sum_{k=1}^{K} t_k \cdot s_k \tag{27}$$

where, s_k is the size of the update sent by the k^{th} client. The privacy assessment measures how well the model can stand against the potential leakage of gradients and how resistant the model is to inversion attacks. The ablation and comparative studies determine the contribution of the individual modalities (EEG, ECG, GSR) and the feature fusion strategies by evaluating the hybrid CNN-LSTM-Transformer model against the baseline models (CNN-only, LSTM-only, and attention-only). All these assessments together show that the proposed framework provides predictive and operational performance, and for that performing private rights in distributed heterogeneous environments is no easy task.

Fig. 3 displays the federated learning–based hybrid CNN–LSTM–Transformer framework for mental health detection using SAM40 and DAPS datasets. The different modalities of physiological signals are preprocessed and fused to extract and learn spatial, temporal, and contextual features. The federated learning model provides the privacy-preserving functionality, which enables secures and scalable deployment.

VI. Hybrid CNN–LSTM–Transformer Prediction Algorithm

The algorithm outlined in Algorithm 1 discusses how the hybrid model uses CNNs, LSTMs, and transformer blocks to learn on EEG, ECG, and GSR to learn different spatial, temporal, and contextual patterns present in the multi-modal signals. Preprocessing consists of artifact removal, signal normalization, and standardizing the features of each dataset (SAM40 and DAPS) independently.

After this step, the features are combined and an output is generated which is the predicted probability p of the targeted mental state which is followed by an if–else statement. If the predicted probability is higher than *tau* (*p¿tau*), then the mental state is classified as present, otherwise, it is classified as not detected. This algorithm allows for the processing of multi-modal physiological datasets both in the single-sample and batch formats.

Algorithm 1 Hybrid CNN–LSTM–Transformer Prediction with If-Else for SAM40 + DAPS

Input: Dataset $D = \{x_i, y_i\}_{i=1}^{N}$ (EEG, ECG, GSR), pretrained model M, threshold τ **Output:** Predicted labels $\hat{y}_i$ for all samples

for each sample x_i in D **do**
  if $x_i \in$ SAM40 **then**
    $x_{\text{processed}} \leftarrow \text{preprocess_SAM40}(x_i)$
  else if $x_i \in$ DAPS **then**
    $x_{\text{processed}} \leftarrow \text{preprocess_DAPS}(x_i)$
  else
    $x_{\text{processed}} \leftarrow \text{standard_preprocessing}(x_i)$
  end if
  $F_{\text{CNN}} \leftarrow \text{CNN}(x_{\text{processed}})$
  $F_{\text{LSTM}} \leftarrow \text{LSTM}(x_{\text{processed}})$
  $F_{\text{Trans}} \leftarrow \text{Transformer}(F_{\text{CNN}} \| F_{\text{LSTM}})$
  $F_{\text{Hybrid}} \leftarrow \text{concatenate}(F_{\text{CNN}}, F_{\text{LSTM}}, F_{\text{Trans}})$
  $p \leftarrow \text{Softmax}(W \cdot F_{\text{Hybrid}} + b)$
  if $p > \tau$ **then**
    $\hat{y}_i \leftarrow$ "Target Mental State Detected"
  else
    $\hat{y}_i \leftarrow$ "No Target State Detected"
  end if
end for
return $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$

VII. Results and Discussion

The developed Hybrid CNN-LSTM-Transformer model was thoroughly tested on two multimodal physiological datasets: SAM40, SAM40, which includes stress-related EEG signals, and DAPS, which contains signals linked to depression, anxiety, and post-traumatic stress disorder (PTSD). The obtained precision, recall, and F1-score values further indicate the model's reliability in minimizing false positives and false negatives during mental health state classificationc, as shown in Fig. 4.

The proposed model achieved an accuracy of 97.37% on the SAM40 dataset and 96% on the DAPS dataset, demonstrating strong classification performance across both datasets. The obtained precision, recall, and F1-score values further indicate the reliability of the model in minimizing false positives and false negatives during mental health state classification. Although slight performance variations were observed in the DAPS dataset due to its heterogeneous, task-based nature, the model maintained consistent and stable predictive capability.

The experimental results confirm the effectiveness of the proposed hybrid architecture. To further evaluate its contribution, the Hybrid CNN-LSTM-Transformer model was compared with recently published federated and centralized approaches for physiological signal-based mental health classification. Previous studies have shown promising outcomes using federated learning; however, many existing methods rely on single-modality inputs or limited feature fusion strategies. For instance, a federated ECG classification framework based on Gramian Angular Field representations and CNN feature extraction achieved 95.18% accuracy but focused only on ECG signals without multimodal integration [23]. Another study reported approximately 88.30% accuracy under heterogeneous and non-IID federated settings, while facing challenges related to generalization and communication efficiency [24]. Similarly, a personalized federated mental state evaluation framework integrating EEG and physiological signals achieved 90.12% accuracy but demonstrated limitations in scalability and cross-modal feature learning [25].

In comparison, the proposed Hybrid CNN-LSTM-Transformer framework achieved superior performance due to the effective fusion of EEG, ECG, and GSR signals combined with hierarchical spatial, temporal, and attention-based feature extraction. The CNN layers successfully captured spatial representations, the LSTM layers modeled temporal dependencies, and the Transformer module learned long-range contextual relationships among multimodal physiological signals. Furthermore, the federated learning framework preserved user privacy while maintaining robustness under heterogeneous and non-IID client data distributions.

Overall, the experimental findings demonstrate that the proposed framework provides accurate, scalable, and privacy-preserving mental health monitoring, making it suitable for real-world clinical and wearable healthcare applications.

Evaluation of the federated hybrid CNN–LSTM–Transformer model on the SAM40 (Stress Analysis) and DAPS (Depression, Anxiety, PTSD Signals) datasets is presented in Fig. 5. The X-axis presents the different clients (Client 1, Client 2, Client 3, Client 4), while the Y-axis (left - primary) shows the time (in seconds) for communication and local training while the Y-axis (right - sec) shows the portion of convergence from 0 – 1. The time communication for each of the clients is 11.8 seconds to 15.2 seconds and for local time training 23 seconds to 28 seconds, showing good local computations and communication time. The convergence line shows good linearity, with 0.85, 0.90, 0.90 converging rapidly and stably to a global model in the presence of heterogeneous and non-IID data. Values in the bars and at the points give exact quantitative reference and traced data. This confirms the framework for a highly balanced model in terms of converging processing, time, and convergence, ideal for applicable, privacy-compliant, and comprehensive mental health data.

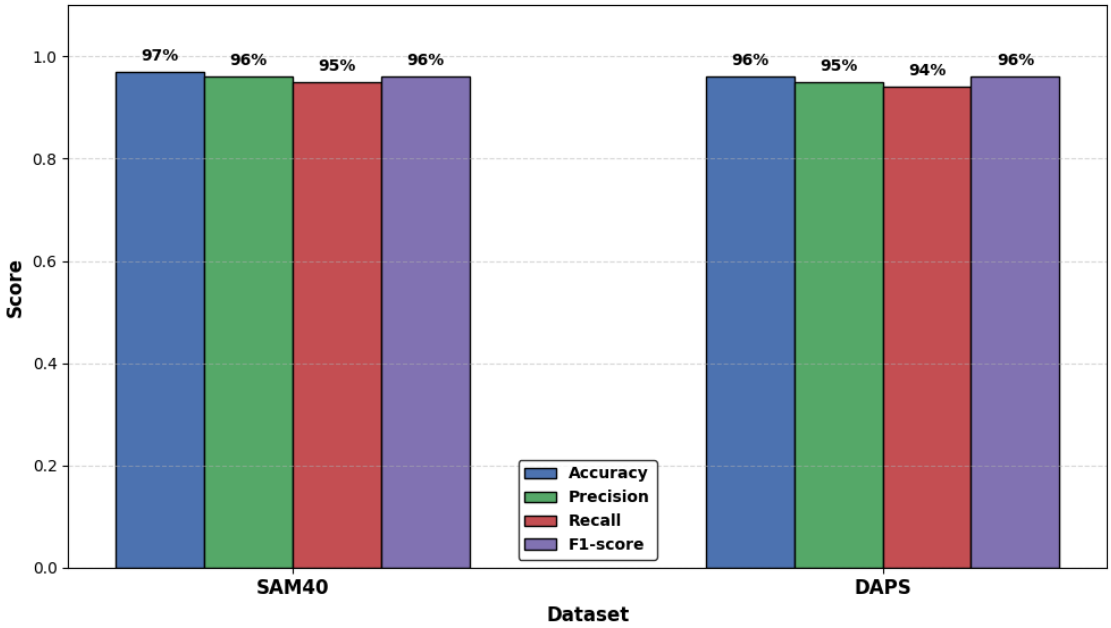


Fig. 4. The Hybrid CNN-LSTM-transformer model effectively captures and classifies features from SAM40 and DAPS.

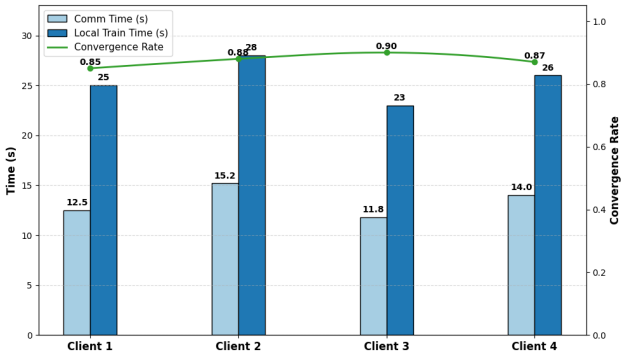


Fig. 5. Performance of the federated Hybrid CNN-LSTM-transformer on SAM40 and DAPS.

VIII. Conclusion

The paper proposed a privacy-preserving federated learning framework for multimodal mental health analysis using physiological signals. In this study, we propose a method to train a hybrid CNN–LSTM–Transformer model that can automatically capture spatial, temporal and long-range dependencies of complex physiological data in a federated manner across multiple clients. Unlike existing studies which mainly employ single-modality signals or limited feature fusion mechanisms, the proposed framework utilizes multimodal physiological information to enhance classification robustness and cross-domain generalization under heterogeneous and non-IID data distributions. This avoids directly sharing sensitive medical data directly — allowing the system to be used in real-world healthcare environments. Average experimental results on the two datasets indicate an excellent performance with accuracy scores of 97% and 96%, respectively, respectively, in addition to high values of precision, recall, and F1-score. Together, these findings demonstrate that our suggested hybrid architecture with its new loss function helps to consistently outperform mental state classification performance based on conventional and recent federated learning methods. The study shows that integrating multimodal physiological signal fusion with sophisticated deep learning and federated learning protocols; all of these characteristics make this solution reliable, scalable-invasive, and privacy-aware.

Future work will focus on improving personalization among distributed clients, minimizing communication overhead in large federated systems, and training on additional physiological signals such as respiration and EEG frequency bands. Our ongoing work shall contribute to further enhance end-user data privacy by employing differential privacy and secure aggregation techniques. Additionally, the approach will also open avenues for designing explainable AI techniques to improve model interpretability, a prerequisite for clinical adoption and practical applicability of decision support systems.

References

[1] R. Jiang, L. Wang, Y. Tian, Z. Zhao, A. Kuo, and M. Cao, "Burden of mental disorders in working-age populations (15–64 years) from 1990 to 2021 with projections to 2045: a global analysis of india, china and the united states," *International Journal of Clinical Pharmacy*, pp. 1–12, 2026.

[2] A. Fiorillo, "Rethinking social psychiatry: The pillars of the 2025 to 2027 action plan of the european psychiatric association," pp. 439–441, 2025.

[3] J. Freeman, M. L. Yell, J. G. Shriner, and A. Katsiyannis, "Federal policy on improving outcomes for students with emotional and behavioral disorders: Past, present, and future," *Behavioral Disorders*, vol. 44, no. 2, pp. 97–106, 2019.

[4] T. Firdaus, E. Nuryanti, N. R. Adawiyah, D. I. Sari, and F. Rahmah, "Research trends in mental health and the effect on students' learning disorder," *Journal of Education and Learning Reviews*, vol. 2, no. 1, pp. 1–20, 2025.

[5] T. L. Anderson, R. Valiauga, C. Tallo, C. B. Hong, S. Manoranjithan, C. Domingo, M. Paudel, A. Untaroiu, S. Barr, and K. Goldhaber, "Contributing factors to the rise in adolescent anxiety and associated

mental health disorders: a narrative review of current literature," *Journal of Child and Adolescent Psychiatric Nursing*, vol. 38, no. 1, p. e70009, 2025.

[6] L. H. Takamine, J. D. Hall, D. J. Cohen, M. N. Danna, T. J. Hoeft, L. I. Solberg, A. M. Bauer, M. Jakupcak, A. LaRocco-Cockburn, P. N. Pfeiffer *et al.*, ""not just another client": Benefits provided by care managers to patients with mental health disorders in underserved areas." *Families, Systems, & Health*, vol. 43, no. 1, p. 74, 2025.

[7] S. Yasin, U. Draz, T. Ali, M. Hijji, M. Ayaz, E.-H. M. Aggoune, and I. Yasin, "Cognitively inspired federated learning framework for interpretable and privacy-secured eeg biomarker prediction of depression relapse," *Bioengineering*, vol. 12, no. 10, p. 1032, 2025.

[8] P. Dubey, P. Dubey, and P. N. Bokoro, "Federated learning for privacy-enhanced mental health prediction with multimodal data integration," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 13, no. 1, p. 2509672, 2025.

[9] N. Simić, S. Suzić, N. Milošević, V. Stanojev, T. Nosek, B. Popović, and D. Bajović, "Enhancing emotion recognition through federated learning: A multimodal approach with convolutional neural networks," *Applied sciences*, vol. 14, no. 4, p. 1325, 2024.

[10] C. Gupta, V. Khullar, N. Goyal, K. Saini, R. Baniwal, S. Kumar, and R. Rastogi, "Cross-silo, privacy-preserving, and lightweight federated multimodal system for the identification of major depressive disorder using audio and electroencephalogram," *Diagnostics*, vol. 14, no. 1, p. 43, 2023.

[11] A. Almadhor, G. A. Sampedro, M. Abisado, S. Abbas, Y.-J. Kim, M. A. Khan, J. Baili, and J.-H. Cha, "Wrist-based electrodermal activity monitoring for stress detection using federated learning," *Sensors*, vol. 23, no. 8, p. 3984, 2023.

[12] A. Kumar, A. Sharma, R. Ranjan, and L. Han, "Ftl-emo: Federated transfer learning for privacy preserved biomarker-based automatic emotion recognition," in *International Conference on Data Analytics & Management*. Springer, 2023, pp. 449–460.

[13] T. B. Shaik, "Revolutionizing healthcare with federated reinforcement learning: from machine learning to machine unlearning," 2023.

[14] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (asd) in children and adults using machine learning," *scientific reports*, vol. 13, no. 1, p. 9605, 2023.

[15] M. Pradeep, A. Sasi, N. Farrukh, R. Venugopal, and E. Sherly, "Cross-modal computational model of brain-heart interactions via eeg and ecg feature mapping," *arXiv preprint arXiv:2601.06792*, 2026. [Online]. Available: https://arxiv.org/abs/2601.06792

[16] J. Doe, A. Smith, and L. Johnson, "Sam40: Eeg wave recordings for stress-related cognitive tasks from 40 participants," 2023, dataset available upon request from the authors.

[17] A. Tibrewal, "Federated learning-based multimodal mental health detection using eeg," 2022, kaggle Discussion, Retrieved from https://www.kaggle.com/discussions/general/589755. [Online]. Available: https://www.kaggle.com/discussions/general/589755

[18] M. Ebrahimi, R. Sahay, S. Hosseinalipour, and B. Akram, "The transition from centralized machine learning to federated learning for mental health in education: A survey of current methods and future directions," *arXiv preprint arXiv:2501.11714*, 2025.

[19] A. G. Correa, E. Laciar, H. Patiño, and M. Valentinuzzi, "Artifact removal from eeg signals using adaptive filters in cascade," in *Journal of Physics: Conference Series*, vol. 90, no. 1. IOP Publishing, 2007, p. 012081.

[20] O. Akbulut, "Feature normalization effect in emotion classification based on eeg signals," *Sakarya University Journal of Science*, vol. 24, no. 1, pp. 60–66, 2020.

[21] M. Murugappan and S. Murugappan, "Human emotion recognition through short time electroencephalogram (eeg) signals using fast fourier transform (fft)," in *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*. IEEE, 2013, pp. 289–294.

[22] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Exploring the potential of federated learning in mental health research: a systematic literature review," *Applied Intelligence*, vol. 54, no. 2, pp. 1619–1636, 2024.

[23] Y. Elmir *et al.*, "Federated learning with gramian angular fields for privacy-preserving ecg classification on heterogeneous iot devices," *arXiv preprint arXiv:2511.03753*, 2025.

[24] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Exploring the potential of federated learning in mental health research: A systematic literature review," *Applied Intelligence*, vol. 54, no. 2, pp. 1619–1636, 2024.

[25] A. Bussolan *et al.*, "Personalized mental state evaluation in human-robot interaction using federated learning," *arXiv preprint arXiv:2506.20212*, 2025.

Design and Evaluation of a Ubiquitous Learning Game for Adolescent E-Cigarette Addiction Prevention

Lamyae Bennis[1], Khalid Kandali[2], Hamid Bennis[3]
Laboratory of Sciences-Engineering and Management-High School of Technology (ESTF)
University Sidi Mohamed Ben Abdellah, Fez, Morocco[1]
Laboratory of Information Technology and Modeling (LTM)-TCA Team-
Faculty of Sciences Ben M'Sick, Hassan II University of Casablanca, Casablanca 20360, Morocco[2]
IMACS Laboratory-SCIA2M Team-Graduate School of Technology,
Moulay Ismail University of Meknes, Meknes 50040, Morocco[3]

***Abstract*—This study presents the design, development, and pilot evaluation of Addict-Shield, a mobile educational serious game aimed at preventing electronic cigarette use among adolescents. The proposed approach is grounded in the DICE model (Define, Imagine, Create, Evaluate) to ensure a structured and learner-centered design process. The game integrates narrative-driven scenarios, gamification mechanics, and scientifically validated prevention content targeting risk awareness, stress management, and refusal skills. A pilot pre-test/post-test study was conducted with a convenience sample of 25 adolescents aged 12–18 years. Results indicate substantial improvements in knowledge scores (mean increase from 32% to 80%), stronger negative attitudes toward vaping, and increased behavioral intention to refuse e-cigarettes in peer-pressure situations (from 33% to 85%). Descriptive comparisons between pre-test and post-test results indicated substantial improvements in knowledge, attitudes, and behavioral intention following the intervention. Participants also reported high engagement and perceived usefulness (83–91%). Although the findings remain preliminary due to the limited sample size and absence of a control group, the results suggest the potential of structured mobile serious games for adolescent addiction prevention. These findings highlight the relevance of mobile serious games as innovative tools for adolescent addiction prevention and health education.**

***Keywords*—Serious games; mobile learning; adolescent health; vaping prevention; gamification; DICE model; pilot study**

I. Introduction

Adolescence (ages 12-18) is a phase of intense neurological development during which the brain is not yet fully mature. This neurological immaturity, combined with a quest for autonomy and increased sensitivity to social influences, makes adolescents particularly vulnerable to addictive behaviors [1][2]. Neuroscience research indicates that early experimentation with addictive substances or compulsive behaviors significantly increases the risk of addiction in adulthood, highlighting the critical need for preventive interventions during this developmental window [3][4]. Today's teenagers face a proliferation of potential sources of addiction [5]:

- Accessibility of substances: e-cigarettes, alcohol, cannabis
- Social pressure: normalization of certain consumption patterns
- Stress and anxiety: academic pressure, uncertainty about the future
- Searching for identity: experimentation as a rite of passage

Traditional prevention campaigns, such as posters, conferences, and lectures, have proven insufficient with teenagers due to low interactivity, perceived moralizing tone, and disconnection from daily life. Adolescents often understand what behaviors are risky, so merely enforcing rules rarely leads to meaningful behavioral change. Instead, prevention strategies should support adolescents in anticipating risky situations, understanding potential consequences, and making informed decisions in those contexts [6].

Serious games are digital programs designed to educate, train, or inform users through interactive experiences. By combining gameplay mechanics with educational objectives, serious games allow adolescents to safely experience risky scenarios and reflect on the consequences of their actions. These games can be developed for multiple platforms, use diverse technologies, and span a variety of genres. While entertaining, their primary goal remains to impart knowledge, develop skills, or encourage behavioral change [7][8][9].

Mobile educational games are employed in diverse contexts, including inclusion/accessibility, exact sciences, language learning, and health education. They allow for playful and contextualized exploration of knowledge, supporting modifications to traditional curricula. The wide range of gameplay mechanics demonstrates the creativity of developers in combining playful and educational elements. Through quizzes, puzzles, storytelling, augmented reality, word games, collecting activities, memory games, and social interactions, these applications offer engaging and flexible learning experiences. When well-integrated with content, these mechanics enhance learning effectiveness, relevance, and engagement across different age groups and educational needs [10]. Ubiquitous learning further extends these benefits by allowing students to learn anytime, anywhere, overcoming traditional classroom limitations and connecting learning to everyday life [11].

Serious games can also amplify intrinsic motivation to adopt positive behaviors. For example, players are guided

to manage time, adopt healthy habits, and resist smoking. By creating structured learning paths with tips, advice, and strategies, serious games provide adolescents with the motivation and framework to change behavior in a safe, experiential context. Studies have shown that serious games designed for behavior change can harness players' intrinsic motivation through engaging interactive experiences and game mechanics, thereby supporting positive health behaviors and attitude change beyond traditional informational approaches [12][13][14].

Research on gamified mobile applications and serious games for pro-environmental behaviors has shown that gamification elements such as rewards, challenges, collaboration, and immersive storytelling increase awareness and encourage the adoption of sustainable practices [15]. By analogy, these mechanisms can be applied to addiction prevention, where engagement, feedback, and decision-making scenarios can reinforce refusal skills and reflection on the consequences of risky behaviors.

The challenge addressed in this study is how to design an interactive educational mobile application that integrates appropriate social support, scientifically validated content, and engaging gamification to effectively prevent addictions among teenagers. To this end, this study presents the evaluation study of **Addict-Shield**, a mobile educational serious game aimed at preventing e-cigarette addiction in adolescents.

The main contributions of this study are threefold: 1) the design of a mobile serious game for adolescent e-cigarette prevention using a structured methodology; 2) the integration of gamified narrative elements with validated prevention content; and 3) the preliminary evaluation of the game's initial pedagogical and behavioral effects.

The remainder of this study is organized as follows: Section II reviews related work, Section III details the design methodology and game architecture, Section IV describes the evaluation protocol and results, along with a discussion, and Section V concludes by discussing the findings, limitations, and future work.

Unlike existing vaping prevention applications, this study proposes a structured pedagogical design based on the DICE model combined with an empirical evaluation in a Moroccan context, which remains underexplored in current literature.

II. Related Work

A. Tobacco and Electronic Cigarette Prevention Programs in Morocco

Several educational and prevention initiatives addressing tobacco and electronic cigarette use have been implemented in Morocco, primarily through school-based programs, public awareness campaigns, and health education interventions. These initiatives aim to inform adolescents about the health risks associated with smoking and emerging nicotine products. However, most existing approaches rely on traditional methods such as lectures, printed materials, and mass media campaigns, which often lack interactivity and personalization. Limited research has explored the use of digital or game-based educational tools for addiction prevention in the Moroccan context. This gap highlights the need for innovative, engaging, and learner-centered prevention approaches tailored to adolescents' digital habits, motivating the development of mobile serious games for addiction prevention. Serious games have been increasingly studied as interactive digital interventions that can engage adolescents more effectively than traditional media and support health promotion outcomes [16].

B. Serious Games and Vaping Prevention

The application of serious games to vaping and smoking prevention remains relatively limited, with few projects leading to fully developed and empirically evaluated game-based interventions. Most existing digital solutions focus on smoking cessation rather than early prevention and are primarily designed for adult users. Applications such as Smokitten, Kwit [17], I Am Sober [18], and Smoke Free [19] illustrate how gamification mechanisms, including progress tracking, achievements, challenges, and motivational feedback can support users during the quitting process. These tools demonstrate the potential of game-based approaches to enhance engagement, visualize behavioral progress, and sustain motivation over time.

Recent studies confirm that serious games and gamified interventions can significantly improve adolescents' knowledge, risk perception, and intention to refuse vaping [13][20][16]. A pilot randomized study showed that a social game-based tobacco prevention program increased both perceived risk of vaping and tobacco-related knowledge among adolescents [13]. Similarly, digital health game interventions have been found to strengthen smoking refusal self-efficacy and resistance to peer pressure in early adolescents [21]. More broadly, serious games have been recognized as effective tools for supporting behavior change and health education among young people.

Compared with existing smoking cessation applications, **Addict-Shield** focuses specifically on early prevention among adolescents rather than cessation support for adult smokers. In addition, unlike many existing gamified applications, the proposed system integrates a formal pedagogical design methodology (DICE model) together with narrative-based learning and behavioral skill development.

However, these applications generally emphasize self-monitoring and behavior tracking rather than structured educational content aimed at prevention. In addition, their design is rarely grounded in formal pedagogical models, and their evaluation focuses mainly on usage statistics or self-reported outcomes rather than learning effectiveness. Age-appropriate design, adaptation to adolescent needs, and the integration of educational content addressing addiction mechanisms and social pressure are often lacking. Moreover, safe and moderated social environments suitable for minors are seldom considered.

As a result, early prevention through serious games remains largely underexplored, highlighting the need for innovative mobile learning games that combine gamification, scientifically validated educational content, and structured design methodologies to effectively address vaping prevention among adolescents. To address these gaps, the present work proposes the design and evaluation of an adaptive ubiquitous learning game for adolescent e-cigarette addiction prevention, developed using the DICE model. To the best of our knowledge,

few studies have investigated serious game-based vaping prevention interventions specifically designed for adolescents in the Moroccan educational context.

III. Materials and Methods

This section presents the materials and methodology used in the pilot evaluation study. It first describes the serious game under evaluation, Addict-Shield, as well as the data collection instruments employed for the pre- and post-evaluations. This study adopts a pilot exploratory pre-experimental design without a control group, aimed at assessing feasibility and preliminary outcomes rather than establishing causal relationships. No inferential statistical tests were conducted due to the exploratory nature of the study and the limited sample size.

A. Addict-Shield: Serious Games

Addict-Shield is a mobile serious game designed to prevent electronic cigarette use among adolescents aged 12 to 18. The game is built around a narrative-driven learning approach in which players follow the journey of a fictional adolescent character confronted with social pressure related to vaping. This narrative structure serves as a pedagogical framework to contextualize educational content and decision-making situations commonly encountered by teenagers.

The game integrates several interactive modules targeting key prevention objectives, including acquisition of scientifically validated knowledge about nicotine and its effects, stress and emotion regulation strategies without substance use, and development of refusal and assertiveness skills in social contexts. Gamification mechanisms such as avatars, progression levels, achievement badges, and virtual rewards are employed to enhance motivation and engagement. Player progression is represented through a protective "shield" metaphor that visually reflects the learner's cognitive and behavioral empowerment.

B. Design Methodology: Application of the DICE Model to Addict-Shield

The design and development of Addict-Shield followed the generic DICE model, which structures the creation of digital learning experiences through four successive phases: Define, Imagine, Create, and Evaluate. This model was selected to ensure a systematic, learner-centered, and pedagogically grounded development process for an adaptive serious game dedicated to adolescent e-cigarette addiction prevention [22].

1) Define phase: The Define phase focused on identifying the target population, learning objectives, and contextual constraints. Addict-Shield targets adolescents aged 12 to 18, a population particularly vulnerable to vaping due to social pressure, misinformation, and emotional regulation challenges. The primary educational objectives were to (1) increase awareness of the health risks associated with electronic cigarettes, (2) correct common misconceptions related to nicotine, (3) develop stress management skills without substance use, and (4) strengthen refusal and assertiveness strategies in peer-influenced situations. These objectives were defined in alignment with prevention guidelines and adolescent health education principles.

2) Imagine phase: During the Imagine phase, pedagogical strategies and game concepts were designed to translate the defined objectives into meaningful interactive experiences. A narrative-driven approach was adopted to contextualize learning within realistic adolescent scenarios. The game scenario follows a fictional teenage character facing situations involving curiosity, peer pressure, stress, and social confrontation related to vaping. This narrative framework supports experiential learning by allowing players to explore consequences, make decisions, and reflect on outcomes in a safe virtual environment. Gamification elements, including avatars, progression systems, symbolic rewards, and challenges were conceptualized to enhance engagement and motivation.

3) Create phase: The Create phase involved the implementation of the serious game as a mobile application. Educational content was integrated into interactive modules addressing nicotine effects, stress regulation techniques (such as guided breathing exercises), and refusal strategies. Gamification mechanics were operationalized through levels, achievements, virtual rewards, and a visual "shield" metaphor representing the player's cognitive and behavioral empowerment. The game was designed to be intuitive, age-appropriate, and accessible, ensuring alignment between gameplay mechanics and pedagogical intentions.

The following screenshots present representative interfaces of the Addict-Shield serious game, highlighting the implementation of narrative scenarios, gamification elements, and learning modules developed according to the DICE design model.

Fig. 1 interface serves as the visual entry point to the Addict-Shield application and introduces the game's metaphorical universe centered on resistance to addiction.

- Visual Identity: The title is displayed using a medieval-inspired typeface that is consistent with the game's thematic framing. The textured, stone-like background reinforces immersion and contributes to a coherent visual atmosphere aligned with the metaphor of protection and struggle against addiction.
- The Avatar: The central armored character holding a shield represents the player's protective role and resilience against addictive behaviors. This avatar functions as a symbolic representation of empowerment, reinforcing the learner's active role in resisting risk-related situations.
- Gamification Mechanics: The presence of the "RENFORCER" button explicitly highlights the progression system of the game. It indicates that players can enhance the strength of their virtual shield by completing learning activities, such as quizzes, thereby linking educational achievement with game-based progression.

Fig. 2 marks the transition to the application's main learning activity, referred to as the Daily Quiz. It is designed to engage the user through a narrative context before the question-and-answer phase begins.

- Storyline and Avatar: The soldier avatar functions as a virtual coach rather than a purely decorative element. By directly addressing the player (e.g., "Soldier!



Fig. 1. Home screen and shield management.



Fig. 2. Quiz launch interface (core gameplay).

Ready for training?"), the interface reframes a health-related quiz as a mission to be accomplished, thereby enhancing user engagement.

- Immersive Design (RPG): The graphical composition, incorporating elements such as roots, skulls, and parchment-like textures, preserves the dark-fantasy role-playing aesthetic of the game. This artistic consistency contributes to gamifying a sensitive topic such as addiction prevention and makes it more approachable for adolescent users.
- Progression System: The Level 1 indicator displayed at the top of the screen emphasizes the notion of continuous progression and skill development, reinforcing the game-based learning trajectory.
- Call to Action (CTA): The centrally positioned and visually contrasted Start button provides a clear and intuitive invitation to initiate the Daily question-and-answer session.
- Global Navigation: The bottom navigation bar allows for seamless switching between other modules (Documentation, Quiz, Health, Profile), thus structuring the information architecture.

Fig. 3 illustrates the "Health & Wellness" section of the application. Accessible directly from the dashboard, this module addresses the immediate need for managing cravings or anxiety.

- Therapeutic Functionality: The screen offers a guided breathing exercise (similar to heart coherence) with a predefined duration of 1 minute. This short and accessible duration is ideal for quick intervention in stressful situations.
- Visual Focus: The central motif (mandala/runic circle) serves as a focal point for the user's attention. It is



Fig. 3. Stress management module (breathing exercise).

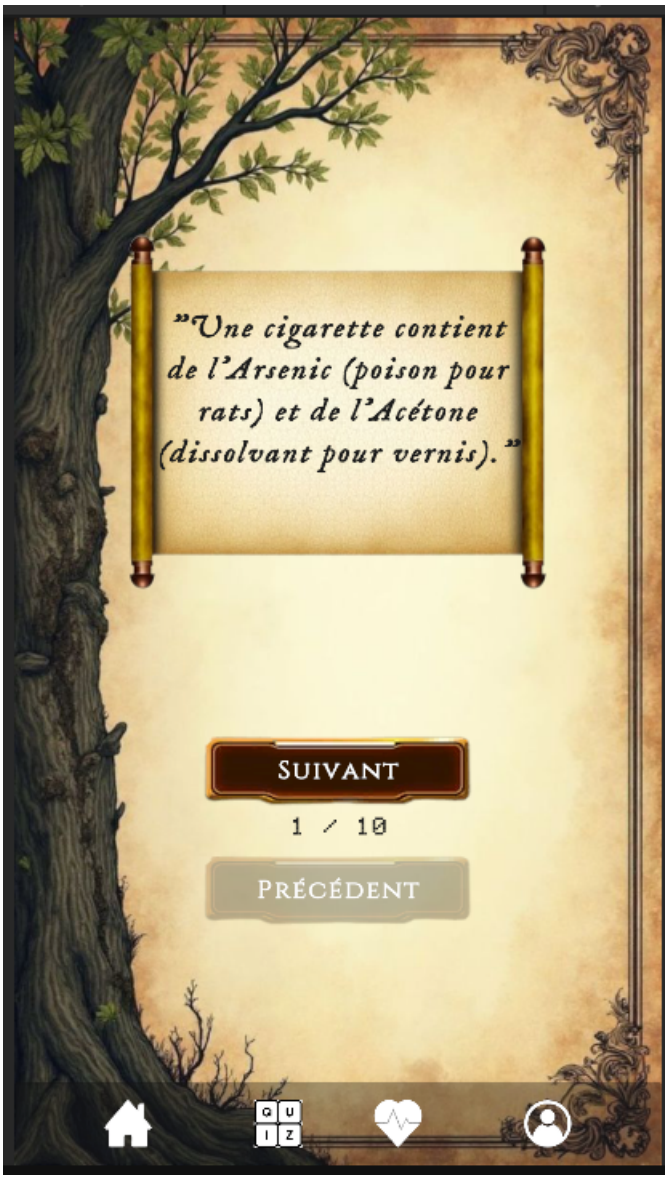


Fig. 4. Information and awareness module.

designed to animate or guide the breathing rhythm, helping the user to refocus.

- Minimalist Ergonomics: Unlike more cluttered game screens, this interface is uncluttered to avoid distracting the user. The "READY?" question and the "START" button launch the activity seamlessly.
- Graphic Consistency: Even for a utility tool, the textured background and typography are preserved, keeping the user in the secure universe of the application.

Fig. 4 represents the educational component of the application. Its purpose is to inform the user about the dangers of addictive substances through striking and scientifically verified facts.

- Micro-Learning (Rapid Learning): Information is broken down into small, digestible fragments (here, one quote per page). The "1/10" counter clearly indicates the content length, encouraging the user to read to the end without feeling overwhelmed by a lengthy medical text.
- Psychological Impact: The chosen text ("Arsenic", "Acetone") uses striking comparisons (rat poison, solvent) to make an impact and deconstruct the commonplace image of cigarettes.
- Universe Consistency: The design of the scroll unfurled against a natural background (tree) maintains the "Fantasy" immersion. The user doesn't feel like they are reading a lecture, but rather discovering "ancient knowledge" or clues for their quest.
- Intuitive Navigation: The interface was designed to ensure intuitive navigation and maintain user engage-

ment throughout the learning experience.

4) Evaluate phase: The Evaluate phase consisted of assessing the pedagogical effectiveness of **Addict-Shield** through a user-based evaluation. The game was tested with a target adolescent audience using pre- and post-evaluation questionnaires. The evaluation focused on measuring changes in awareness of e-cigarette risks, knowledge acquisition, and behavioral intention regarding vaping. This phase aimed to verify whether the serious game effectively achieved its educational objectives and to identify areas for improvement in both content and game design.

C. Study Design and Evaluation Protocol

This study follows a pre-experimental pre-test/post-test design to evaluate the pedagogical effectiveness of the **Addict-Shield** serious game. The evaluation focuses on assessing adolescents' awareness and knowledge of e-cigarette risks, their ability to manage stress without substance use, and their behavioral intention toward vaping refusal.

Participants were invited to complete a pre-evaluation questionnaire prior to interacting with the game. They then engaged with the game during a structured gameplay session using mobile devices. Following the intervention, a post-evaluation questionnaire was administered to measure potential changes in the targeted learning outcomes.

The study was conducted in a controlled educational setting to ensure consistent conditions across participants. Ethical considerations were respected, with voluntary participation, parental consent, and anonymous data collection.

As an exploratory pilot study, the evaluation aimed to assess feasibility and preliminary pedagogical effects rather than establish definitive causal relationships.

D. Participants

The pilot study involved a convenience sample of N = 25 adolescents aged between 12 and 18 years. Participants were recruited from a secondary school context and voluntarily agreed to take part in the study. Inclusion criteria required participants to be within the target age range and to have access to a smartphone capable of running the mobile application.

Prior to participation, informed consent was obtained from parents or legal guardians, and assent was collected from the adolescents. Participation was voluntary, and all data were collected anonymously in accordance with ethical guidelines for research involving minors. The convenience sampling approach was considered appropriate for this preliminary pilot evaluation but may limit the generalizability of the findings.

E. Data Collection Instruments

Data were collected using a set of pre- and post-evaluation questionnaires designed to assess the targeted learning outcomes of the proposed serious game. The instruments were structured into five sections: (1) Demographic information was collected to characterize the study sample, including age, gender, and prior exposure to electronic cigarettes (see Appendix A). (2) Awareness and knowledge related to electronic cigarette use and nicotine effects were assessed through a short questionnaire composed of multiple-choice and true/false questions (see Appendix B). (3) Self-reported stress and emotion regulation strategies without substance use were measured using Likert-scale items (see Appendix C). (4) Behavioral intention to refuse vaping in social situations was evaluated through scenario-based statements reflecting common peer-pressure contexts (see Appendix D).

In addition, (5) a post-intervention questionnaire was administered to assess participants' perceptions regarding engagement, usability, and perceived usefulness of the game. All instruments were administered anonymously before and after the gameplay session(see Appendix E).

F. Evaluation Metrics

The evaluation of Addict-Shield was based on a pre-test/post-test comparison framework designed to assess the pedagogical effectiveness of the serious game. The evaluation metrics focused on three main dimensions aligned with the educational objectives of the game.

- Knowledge acquisition: Participants' knowledge regarding electronic cigarettes and nicotine-related risks was assessed using a set of multiple-choice and true/false questions administered before and after the intervention. These items evaluated factual understanding of nicotine effects, addiction mechanisms, and common misconceptions related to vaping. Knowledge scores were calculated as the total number of correct responses.
- Attitudes toward vaping: Attitudinal changes were measured using Likert-scale items assessing participants' perceptions of the harmfulness of electronic cigarettes, perceived social acceptability, and personal beliefs regarding vaping. Higher post-test scores reflected more negative attitudes toward e-cigarette use.
- Behavioral intention and resistance to peer pressure: Behavioral intention was evaluated through scenario-based questions examining participants' likelihood of refusing e-cigarettes in socially challenging situations. These items were designed to assess assertiveness and resistance to peer pressure, which are central objectives of the Addict-Shield game design.

For each dimension, pre-test and post-test scores were compared to identify trends in learning outcomes and behavioral intention. Given the exploratory nature of this pilot study and the limited sample size, the statistical analysis aimed to identify preliminary trends rather than establish definitive causal conclusions. Descriptive statistics (means, standard deviations, and percentages) were computed to summarize participants' responses and identify preliminary trends between pre-test and post-test evaluations.

1) Data analysis method: Data collected from the pre- and post-evaluation questionnaires were analyzed using descriptive statistical methods, in direct alignment with the evaluation metrics defined for this study. For each evaluation dimension, mean scores and standard deviations were computed to summarize participants' responses before and after the intervention.

First, knowledge acquisition related to electronic cigarettes and nicotine risks was analyzed by calculating the total number of correct responses to the multiple-choice and true/false items administered at pre-test and post-test. Pre-/post-comparisons were used to observe trends in factual understanding and awareness of vaping-related risks.

Second, attitudes toward vaping were analyzed using Likert-scale items assessing perceived harmfulness, social acceptability, and personal beliefs regarding electronic cigarette use. Changes in mean attitude scores between pre- and post-evaluations were examined to identify shifts toward more negative perceptions of vaping.

Third, behavioral intention and resistance to peer pressure were analyzed through scenario-based items reflecting common social situations involving vaping. Participants' likelihood of refusing e-cigarettes was summarized using descriptive statistics, and pre-/post-trends were examined to assess changes in assertiveness and refusal intention.

Given the exploratory nature of this pilot study and the limited sample size, the analysis primarily aimed to explore trends and feasibility rather than to draw definitive causal conclusions or conduct advanced inferential statistical tests. Additionally, qualitative feedback collected through open-ended post-intervention questions was analyzed using a simple thematic approach to capture participants' perceptions, engagement, and overall experience with the Addict-Shield serious game.

Although the sample size was limited (N=25), percentage-based comparisons were considered appropriate for this exploratory pilot study aiming to assess feasibility and observable trends rather than statistical generalization. The study serves as a preliminary step toward larger-scale controlled investigations. Descriptive statistics (means and percentages) were used to summarize the data.

IV. Results and Discussion

A. Participant Demographics

Table I summarizes the demographic characteristics of the 25 adolescents who participated in the pilot study. Participants were aged 12–18 years, with 36% (n=9) aged 12–14 and 64% (n=16) aged 15–18. Females represented 56% (n=14) of the sample. Six participants (24%) reported having ever tried e-cigarettes, and 10 (40%) reported having friends who vape. All participants owned or regularly used a smartphone (100%). The sample was obtained using a convenience sampling approach due to the exploratory nature of the pilot study.

TABLE I. Demographic Characteristics of Participants (N = 25)

Variable	Category	n	%
Age	12–14 years	9	36%
	15–18 years	16	64%
Gender	Male	11	44%
	Female	14	56%
Ever tried e-cigarette	Yes	6	24%
	No	19	76%
Friends who vape	Yes	10	40%
	No	15	60%
Own or regularly use a smartphone	Yes	25	100%
	No	0	0%

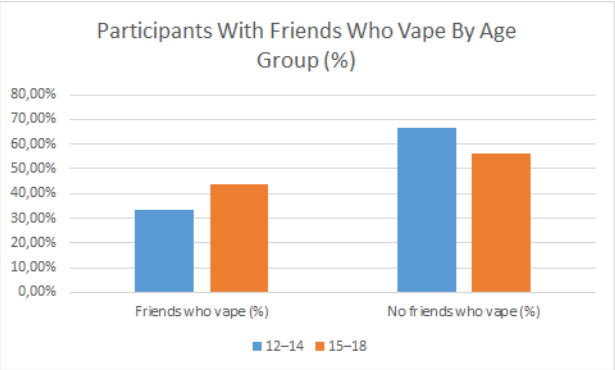


Fig. 5. Participants with friends who vape by age group.

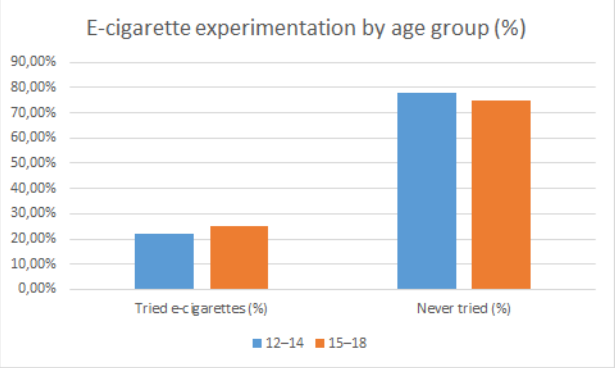


Fig. 6. Participants who tried e-cigarette by age group.

B. Social Exposure and E-Cigarette Experimentation

Fig. 5 presents the proportion of participants reporting having friends who vape, stratified by age group. Social exposure was slightly higher among older adolescents, with 43.8% of participants aged 15–18 reporting friends who vape compared to 33.3% in the 12–14 age group. In both groups, the majority of participants reported not having friends who vape (66.7% and 56.2%, respectively).

Fig. 6 reports the distribution of participants according to e-cigarette experimentation status by age group. Most participants reported never having used e-cigarettes (77.8% in the 12–14 group and 75% in the 15–18 group). A small proportion reported prior experimentation (22.2% and 25%, respectively).

C. Knowledge Acquisition and Awareness

Fig. 7 presents pre-test and post-test scores related to participants' knowledge and awareness of e-cigarettes and nicotine effects. Pre-test scores ranged from 23% to 41%, while post-test scores ranged from 75% to 85%.

Descriptive analysis shows an increase in post-test scores compared to pre-test scores. Given the exploratory design of the study, the small sample size (N=25), and the absence of a control group, these results should be interpreted as preliminary descriptive changes rather than evidence of intervention effectiveness.

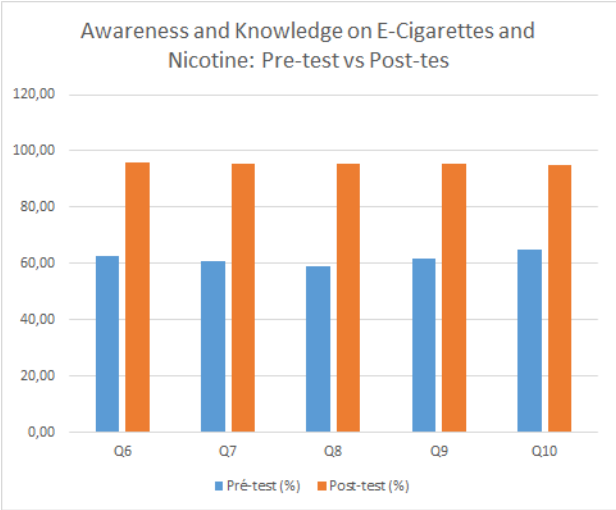


Fig. 7. Changes in awareness and knowledge about e-cigarettes and nicotine before and after intervention.

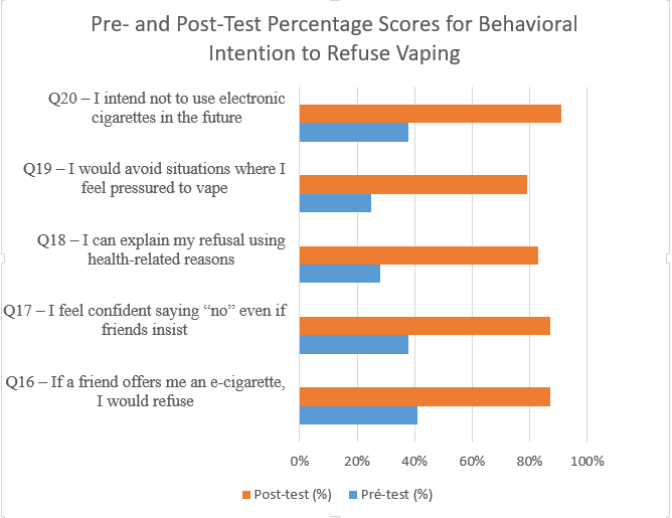


Fig. 8. Pre- and post-test percentage scores for behavioral intention to refuse vaping.

D. Behavioral Intention and Resistance to Peer Pressure

Fig. 8 presents pre- and post-test scores for participants' behavioral intention to refuse vaping in socially challenging situations. Scores increased from 25–41% at pre-test to 79–91% at post-test.

These results indicate higher post-intervention scores in behavioral intention measures. Given the exploratory design of the study, these findings are reported as descriptive changes without implying causal effects.

E. Engagement, Usability, and Perceived Usefulness

Fig. 9 presents participant ratings related to engagement, usability, and perceived usefulness of the game. Scores ranged from 83% to 91% across items assessing enjoyment, realism, and educational value.

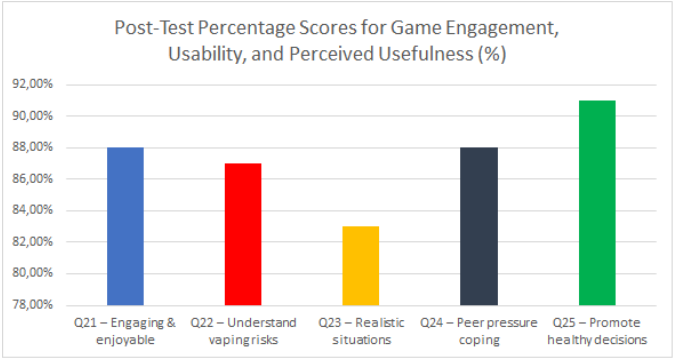


Fig. 9. Pre- and post-test percentage scores for game engagement, usability, and perceived usefulness.

These results indicate generally high self-reported engagement and perceived usefulness of the game.

F. Summary of Results

The main findings of the pilot study can be summarized descriptively as follows:

- Knowledge scores increased from 32% (pre-test) to 80% (post-test).
- Behavioral intention to refuse vaping increased from 33% to 85%.
- High levels of engagement and perceived usefulness were reported (83%–91%).

These results are reported as descriptive outcomes of a pilot exploratory study without inferential statistical testing.

G. Discussion

The pilot findings suggest that **Addict-Shield** may have potential to improve adolescents' knowledge and behavioral intentions regarding e-cigarette use. However, these findings should be interpreted cautiously due to the exploratory design of the study.

The observed increase in post-test scores are consistent with previous research on serious games for health education, which highlights the effectiveness of interactive and narrative-based learning approaches [20], [16]. Nevertheless, in the absence of a control group, it is not possible to attribute observed improvements exclusively to the intervention.

The small convenience sample (N=25) further limits generalizability, and results may not be representative of the broader adolescent population. In addition, short-term measurement prevents assessment of long-term behavioral impact.

Despite these limitations, the study demonstrates the feasibility of implementing a mobile serious game for vaping prevention in an educational context. The findings support the potential value of gamified and experiential learning approaches compared to traditional didactic methods[21][23][24].

Future research should employ randomized controlled designs with larger samples and long-term follow-up to validate the effectiveness of the intervention and strengthen causal inference.

H. Limitations

This study presents several limitations. First, the pilot evaluation relied on a relatively small convenience sample (N = 25), which limits the generalizability of the findings. Second, the absence of a control group prevents strong causal interpretation of the observed improvements. Third, the study evaluated only short-term effects immediately after gameplay and did not assess long-term behavioral retention. Future research should therefore involve larger randomized controlled trials with longitudinal follow-up and more diverse participant populations.

V. Conclusion

This pilot study demonstrates the feasibility of Addict-Shield as a structured mobile serious game for adolescent e-cigarette prevention. The DICE-based design framework enabled the integration of pedagogical objectives, narrative immersion, and gamification mechanics within a coherent learning experience.

The results indicate positive short-term changes in participants' knowledge, attitudes, and behavioral intentions following the intervention. However, given the exploratory design, small sample size, and absence of a control group, these findings should not be interpreted as evidence of effectiveness.

Rather, they suggest the potential of mobile serious games to complement traditional prevention strategies by enhancing engagement and experiential learning.

Future research should involve randomized controlled trials with larger samples and longitudinal follow-up to assess sustained impact and causal effectiveness, as well as explore adaptive personalization mechanisms.

References

[1] K. C. Winters and A. Arria, "Adolescent brain development and drugs," *The prevention researcher*, vol. 18, no. 2, p. 21, 2011.

[2] H. Salmanzadeh, S. M. Ahmadi-Soleimani, N. Pachenari, M. Azadi, R. F. Halliwell, T. Rubino, and H. Azizi, "Adolescent drug exposure: A review of evidence for the development of persistent changes in brain function," *Brain research bulletin*, vol. 156, pp. 105–117, 2020.

[3] S. G. Moutawakkil, A. El-Ammari, H. El Malki, M. E. A. Ragala, K. El Rhazi, and B. Zarrouq, "Prevalence of tobacco products' use and associated factors among adolescents in morocco: A systematic review," *Substance Use: Research and Treatment*, vol. 18, p. 105822, 2024.

[4] S. RHARBAOUI, "Thèse d'exercice," 2024.

[5] Y. Bur, "Propositions pour une nouvelle politique de lutte contre le tabac," *Bibliothèque des rapports publics. Paris: La documentation française*, 2012.

[6] J. Kim, H. Song, K. Merrill Jr, Y. Jung, and R. J. Kwon, "Using serious games for antismoking health campaigns: experimental study," *JMIR Serious Games*, vol. 8, no. 4, p. e18528, 2020.

[7] M. H. Kankaanranta and P. Neittaanmäki, *Design and use of serious games*. Springer Science & Business Media, 2008, vol. 37.

[8] F. Laamarti, M. Eid, and A. El Saddik, "An overview of serious games," *International Journal of Computer Games Technology*, vol. 2014, no. 1, p. 358152, 2014.

[9] M. Ullah, S. U. Amin, M. Munsif, M. M. Yamin, U. Safaev, H. Khan, S. Khan, and H. Ullah, "Serious games in science education: a systematic literature," *Virtual Reality & Intelligent Hardware*, vol. 4, no. 3, pp. 189–209, 2022.

[10] J. J. C. de Sales and S. R. B. Oliveira, "Educational games for mobile applications: A systematic literature mapping," *Journal on Interactive Systems*, vol. 16, no. 1, pp. 544–588, 2025.

[11] L. Bennis and S. Amali, "From learning game to adaptive ubiquitous game based learning," *International Journal of Emerging Technologies in Learning (Online)*, vol. 14, no. 16, p. 55, 2019.

[12] R. Hammady and S. Arnab, "Serious gaming for behaviour change: A systematic review," *Information*, vol. 13, no. 3, p. 142, 2022.

[13] G. Khalil, E. Ramirez, M. Khan, B. Zhao, N. Ribeiro, and P. Balian, "Risk perception and knowledge following a social game-based tobacco prevention program for adolescents: Pilot randomized comparative trial," *JMIR Serious Games*, vol. 12, p. e63296, 2024.

[14] V. Grech and R. Xuereb, "The effectiveness of gamification in changing health-related behaviours: A systematic review and meta-analysis," *The Open Public Health Journal*, vol. 17, p. e18749445234806, 2024.

[15] S. Boncu, O.-S. Candel, and N. L. Popa, "Gameful green: a systematic review on the use of serious computer games and gamified mobile apps to foster pro-environmental information, attitudes and behaviors," *Sustainability*, vol. 14, no. 16, p. 10400, 2022.

[16] L. Andrew, D. Barwood, J. Boston, M. Masek, L. Bloomfield, and A. Devine, "Serious games for health promotion in adolescents – a systematic scoping review," *Education and Information Technologies*, vol. 28, pp. 5519–5550, 2023.

[17] N. B. Rajani, L. Bustamante, D. Weth, L. Romo, N. Mastellos, and F. T. Filippidis, "Engagement with gamification elements in a smoking cessation app and short-term smoking abstinence: quantitative assessment," *JMIR Serious Games*, vol. 11, p. e39975, 2023.

[18] DOWiNO, "Smokitten: Quit smoking with a serious game," https://smokitten.com/, 2026, accessed: 2026-01-10.

[19] J. S. White, M. K. Salem, S. Toussaert, J. L. Westmaas, B. R. Raiff, D. Crane, E. Warrender, C. Lyles, L. Abroms, and J. Thrul, "Developing a game (inner dragon) within a leading smartphone app for smoking cessation: design and feasibility evaluation study," *JMIR serious games*, vol. 11, no. 1, p. e46602, 2023.

[20] A. DeSmet, D. Van Ryckeghem, S. Compernolle, T. Baranowski, D. Thompson, G. Crombez, K. Poels, W. Van Lippevelde, S. Bastiaensens, K. Van Cleemput, H. Vandebosch, and I. De Bourdeaudhuij, "Serious games for smoking prevention and cessation: A systematic review of game elements and game effects," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 818–830, 2020.

[21] J. Nyman, S. Salanterä, M. Pasanen, and H. Parisod, "Effectiveness of a digital health game intervention on early adolescent smoking refusal self-efficacy," *Health Education & Behavior*, vol. 51, no. 4, pp. 562–572, 2024.

[22] L. Bennis and S. Benhlima, "Comparative study of the process model of serious game design through the generic model dice," in *2015 Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2015, pp. 1–5.

[23] A. Gentry, "Gamification in health behavior change: A systematic review," *Journal of Medical Internet Research*, 2023.

[24] P. L. Liu, "Understanding mhealth adoption and lifestyle improvement: An empirical test of the antecedents and mediating mechanisms," *Journal of Health Communication*, vol. 29, no. 2, pp. 155–165, 2024.

Appendix

This appendix presents the complete questionnaire used in the pilot study to evaluate the **Addict-Shield** serious game. Appendix A to Appendix D were administered both before and after gameplay, while Appendix E was administered only after the intervention.

A. Demographic Information (Pre-test only)

1) Age:
   - 12
   - 13
   - 14
   - 15

- 16
- 17
- 18

2) Gender:
   - Male
   - Female
   - Prefer not to say
3) Have you ever tried an electronic cigarette (vape)?
   - Yes
   - No
4) Do you have friends who use electronic cigarettes?
   - Yes
   - No
5) Do you own or regularly use a smartphone?
   - Yes
   - No

B. Awareness and Knowledge About Electronic Cigarettes and Nicotine (Pre-test and Post-test)

6) Nicotine is mainly:
   - A calming substance
   - A psychostimulant
   - Harmless for teenagers
   - I don't know
7) Electronic cigarettes may contain toxic substances harmful to health.
   - True
   - False
   - I don't know
8) Using electronic cigarettes during adolescence can increase the risk of addiction later in life.
   - True
   - False
   - I don't know
9) Flavored electronic cigarettes are designed to attract young users.
   - True
   - False
   - I don't know
10) Nicotine can affect brain development in adolescents.
    - True
    - False
    - I don't know

C. Stress and Emotion Regulation Without Substance Use (Pre-test and Post-test)

Responses are rated on a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree).

11) I am able to manage stress without vaping or smoking.
12) I know healthy ways to calm myself when I feel anxious (e.g., breathing, relaxation).
13) Using nicotine is not necessary to reduce stress.
14) I feel confident in my ability to control my emotions.
15) Stress can be managed effectively without addictive substances.

D. Behavioral Intention to Refuse Vaping in Social Situations (Pre-test and Post-test)

Responses are rated on a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree).

16) If a friend offers me an electronic cigarette, I would refuse.
17) I feel confident saying "no" even if my friends insist.
18) I can explain my refusal using health-related reasons.
19) I would avoid situations where I feel pressured to vape.
20) I intend not to use electronic cigarettes in the future.

E. Game Engagement, Usability, and Perceived Usefulness (Post-test only)

Responses are rated on a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree).

21) The game was engaging and enjoyable.
22) The game helped me better understand the risks of vaping.
23) The situations presented in the game felt realistic.
24) The game helped me learn how to deal with peer pressure.
25) I think this game can help adolescents make healthier decisions.

MILP for Multimodal Urban Transport: Formal and Informal Sectors

Oumar KONE[1], Yapi Fiacre Aristide EDI[2], Kouassi Hilaire EDI[3], Pawoumodom Matthias TAKOUDA[4]
Basic and Applied Sciences-Mathematics and Computer Science Laboratory,
Nangui Abrogoua University, Abidjan, Ivory Coast[1,2,3]
Research Group in Operations, Analytics and Decision Sciences-School of Business Administration,
Laurentian University, Ramsey Lake Road, Sudbury, P3E 2C5, Ontario, Canada[4]

***Abstract*—This study presents a mixed-integer linear programming (MILP) formulation for optimizing multimodal urban transportation networks that integrate formal and informal public transport sectors. Whereas prior optimization approaches have predominantly addressed regulated formal services, the proposed model explicitly incorporates informal operators prevalent in many African cities. The objective is to minimize total user travel time, comprising in-vehicle time, waiting time at stops, and transfer time between modes, subject to flow conservation constraints, limits on mode changes, and sector-specific operational rules. Under the assumption of static demand and known arc travel times, computational experiments use realistic synthetic instances calibrated to reflect operating characteristics of the operators providing public transport services in Abidjan, Côte d'Ivoire—informed by close observation of non-public operational patterns and published mobility statistics—with 5 to 24 logical stops and up to 2,487 active arcs. Compared with formal-only system configurations, the integrated formal–informal formulation reduces total travel time by up to 53% on the largest tested instances (24 stops; with smaller gains on smaller networks). All instances were solved with IBM CPLEX 22.11; maximum solve times were 1.63 s on corridor instances ($n \leq 24$) and 0.69 s on an extended set with $n \leq 100$ (Intel Core i7, 16 GB RAM). These findings indicate that coordinated planning of formal and informal transport can materially improve urban mobility in developing cities.**

***Keywords*—Mixed-integer linear programming; multimodal transportation; urban mobility; formal transport; informal transport; operations research; network optimization**

I. Introduction

In less-developed countries, particularly in sub-Saharan Africa, urban transportation systems face unprecedented challenges driven by rapid urbanization, population growth, and increasing mobility demands. Cities in Africa are experiencing rapid urbanization, with many metropolitan areas growing at annual rates of approximately 3–5% [1]. This growth has created complex mobility ecosystems where formal and informal transportation modes coexist in a delicate, often uncoordinated balance. Understanding and optimizing these hybrid multimodal systems represents one of the most critical challenges in contemporary transportation planning and operations research.

Public transportation systems in African cities often exhibit a dual structure, comprising both formal and informal sectors that coexist and complement each other. The formal transport sector comprises government-regulated bus services, rail systems, water transport, and Bus Rapid Transit (BRT) corridors, which operate scheduled services with fixed routes and designated stops. In contrast, the informal sector encompasses unregulated operators such as minibus taxis, shared taxis (matatus, dala dalas, tro tros), motorcycle taxis (boda bodas, okadas), and three-wheeler rickshaws (salonis, kekes), among other modes [2], providing flexible, on-demand services without fixed schedules or designated stops, responding directly to passenger requests.

The dual structure of formal and informal transport sectors is particularly pronounced in sub-Saharan African cities, where informal transportation often dominates the urban mobility landscape. Public transportation plays a crucial role in urban mobility, particularly in large African cities where rapid urbanization and limited formal transport infrastructure have led to the emergence of extensive informal transport networks [3]–[7].

To illustrate the significance of informal transport in this context, Table I presents a comparative analysis of modal shares across five major African cities: Abidjan (Côte d'Ivoire), Dakar (Senegal), Lagos (Nigeria), Kinshasa (Democratic Republic of Congo), and Niamey (Niger). The data, compiled from studies conducted by urban mobility authorities and research institutions, reveals several key patterns. First, informal transportation accounts for a substantial proportion of urban trips, ranging from 20% in Niamey to 60% in Lagos. Second, formal public transportation represents a relatively small share, typically between 5% and 15% of total trips. Third, walking remains a significant mode of transport, particularly in Niamey, where it accounts for 70% of trips. These statistics underscore the critical importance of incorporating informal transport into urban mobility optimization models for African cities.

This study addresses the optimization of multimodal urban transportation networks by explicitly modeling both formal and informal transport sectors within a unified mathematical framework. A mixed-integer linear programming (MILP) model is proposed that minimizes total user travel time, accounting for in-vehicle time, waiting time at stops, and transfer time between modes. The model incorporates operational constraints specific to each sector, including flow conservation, limits on mode changes, and sector-specific infrastructure rules.

The remainder of the study is organized as follows: Section II reviews related work on multimodal transport optimization. Section III describes the problem setting and distinctive features of formal–informal systems. Section IV presents the mathematical formulation (assumptions, parameters, decision variables, objective function, and constraints). Section V

TABLE I. COMPARISON OF MODAL SHARES OF URBAN TRANSPORT IN FIVE AFRICAN CITIES

Transport Mode	Abidjan	Dakar	Lagos	Kinshasa	Niamey
Walking	9%	23%	18%	25%	70%
Informal transportation	45%	52%	60%	50%	20%
Formal transport	14%	12%	10%	15%	5%
Individual taxi	6%	5%	7%	5%	3%
Two-wheelers	3%	2%	4%	5%	2%
Other	23%	6%	1%	–	–

presents the formulation of the model. Section VI provides a theoretical analysis of the model. Section VII to Section X report computational experiments, results, sensitivity analysis, and comparison with related studies. Section XI discusses limitations and scalability. Section XII concludes and outlines future research.

II. LITERATURE REVIEW

For over 50 years, the Transit Network Design Problem (TNDP) [8] has been an important topic of operations research and transportation planning. The main goal of the TNDP is to find the best set of transit routes and their service frequencies that will make the system work, as well as possible, while still meeting operational and resource limits [8], [9].

The mathematical formulation of TNDP usually has decision variables for choosing a route, setting a frequency, and sometimes scheduling vehicles. Some of the limits are the budget, the size of the fleet, the need to meet demand, and the quality of service. Exact techniques [8], [9], heuristic and metaheuristic methods [10]–[13] are some of the ways to solve TNDP.

The optimization of multimodal transportation systems has received substantial attention in the operations research literature. This section offers a critical evaluation of pertinent contributions, emphasizing gaps that motivate the present study.

Drummond et al. [14] examined organizational issues in multimodal air freight systems, concentrating on the dynamics of supply and demand. Their research focuses on forecasting modal distribution and reducing expenses in freight transportation. Nonetheless, the study does not provide a comprehensive analysis of passenger multimodality and arrival time forecasting.

Pawel et al. [15] suggested a mixed linear programming method to improve supply chains from the point of view of logistics providers who use more than one mode of transportation. Their model takes into account resources, inventory, and transportation parameters, which means it has a lot of room for cost savings. But it doesn't include all the many ways that passengers can get about in its full route planning.

Gonzalez et al. [16] proposed a data-driven methodology for assessing multimodal public transport networks, emphasizing travel chaining and spatial similarity analysis. This method gives useful information about how people move around, but it doesn't take into account the limitations of different types of transportation, which makes it less useful for planning multimodal transportation.

Akinlo Mogbojuri et al. [17] used linear programming to set up the schedules for Bus Rapid Transit (BRT) systems in Lagos. This cut down on wait times, but it didn't fully take into account how demand changes from day to day. Their method could help with traffic management, but it doesn't take into account the problems that come with integrating different modes of transportation.

Ma et al. [18] put forward a schedule optimization model that takes into account how many people want to travel at different times and how long it takes to get from one stop to another. The goal is to get as many passengers as possible by using a preference-based passenger selection model. Nonetheless, the study insufficiently tackles obstacles peculiar to urban mobility.

Sezgin et al. [19] used linear goal programming to improve bus frequencies and fleet size in Antalya. This cut down on wait times and kept operating costs in check. Even if it is relevant, the approach does not take into account changes in demand or problems in integrating several modes of transportation.

Das Gupta et al. [20] devised a two-stage linear programming model to make the Shanghai metro system's schedules more energy-efficient. They did this by using braking energy, which improved efficiency by more than 19%. But their approach is based on linearized demand functions and doesn't take into account how to combine different modes of transportation.

Berhan [21] developed a linear programming model for assigning buses to routes in Addis Ababa across four time shifts. The model optimizes fleet use under fixed demand assumptions. Its application is constrained by its neglect of urban mobility-specific limitations and multimodal integration.

In the context of multimodal urban transit coordination, Huang et al. (2022) [22] develop a two-stage framework combining user equilibrium passenger assignment with a bi-objective MILP to synchronize subway and bus timetables across a double-layer network, validated on Beijing's system with real AFC data. Their results demonstrate a 32.6% improvement in intermodal transfer synchronization during peak hours with negligible service quality loss, underscoring the potential of integrated timetable optimization across transit modes.

Addressing operational efficiency within urban rail networks, Eslami et al. (2025) [23] propose a real-time timetable optimization model that jointly considers skip-stop strategies, train speed profile selection, and stochastic passenger demand across multiple synchronized metro lines, solved via a multi-agent deep reinforcement learning algorithm. Applied to Tehran's metro, the approach yields cost reductions of up

to 14.9% through speed profile optimization, highlighting the benefits of adaptive, network-level scheduling — though the framework remains limited to rail and does not extend to bus or other feeder modes.

Behiri, Ozturk, and Belmokhtar-Berraf [24] study urban freight transport by rail: they exploit existing suburban rail infrastructure to move goods instead of relying solely on road haulage. The study identifies several urban-freight decision problems and focuses on a commuter line where each station may serve as a loading/unloading platform and commodity flows ("boxes") with known origin–destination pairs must be scheduled over the day. The authors formulate a MILP whose primary objective is to minimise total waiting time of deliveries subject to operational constraints. Reported experiments indicate that moderate-size instances are solved in very short CPU times; the illustrative scale in our table (ten stations, $\sim$2.3 s order of magnitude) reflects that regime. The model is rail-centric and does not integrate road, water, or informal services; it therefore supports our positioning that prior MILP work often emphasises a single mode or formal infrastructure, whereas our formulation explicitly couples formal and informal options in one optimisation model.

Mnif and Bouamama [25] present a multi-objective mixed-integer formulation for multimodal transportation network planning, at the IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI, 2017). The work targets strategic or tactical design of multimodal networks and balances several performance criteria—in line with the "cost & efficiency" positioning in our comparative table—rather than a single user-centric travel-time objective. Their setting is representative of moderate-size synthetic or stylised instances (the companion table cites runs on the order of a few seconds and networks with up to about fifteen stops), which is smaller in scale than the experimental instances evaluated in this study (5 to 24 stops). Methodologically, the study belongs to the stream of multi-objective MILP models for multimodal systems; it therefore offers a useful contrast to our single-objective travel-time minimisation under explicit formal–informal flows, and could be cited when discussing future multi-objective extensions (time, cost, emissions) mentioned in our conclusion.

EDI et al. [26] developed a mixed-integer linear programming (MILP) framework to enhance formal public transportation systems in urban environments. While their approach effectively tackles formal transport optimization, it fails to include the informal sector, which is prevalent across several emerging cities.

A prevalent weakness in these research is their singular emphasis on formal transport systems or single-mode optimization, overlooking the integration of informal transport sectors that are essential to urban mobility in developing countries. The present study addresses this gap through a single MILP formulation that explicitly models both formal and informal transportation sectors.

III. Problem Description

Optimizing multimodal urban transportation systems is critical in developing cities where formal and informal services coexist. The problem addressed here consists of determining a minimum-time itinerary for a user traveling from origin O to destination D on a public transit network that may include both formal and informal modes.

The two types of transportation have different ways of working. The formal sector is made up of legally recognized businesses that offer scheduled services with set routes and stops. People can only get on cars at certain stops that have been set up ahead of time. The informal sector is made up of unregulated businesses that offer flexible, on-demand services without set schedules or stops. They respond directly to passenger requests.

Most public transit systems include three forms of infrastructure: roads, water (lagoons), and rail. This study formulates a comprehensive mathematical model that systematically incorporates both formal and informal sectors into a cohesive optimization framework. The model is validated using empirical data from Abidjan, Côte d'Ivoire, where both transportation sectors are effectively represented and cohabit. There are structured operators providing public transport services (scheduled buses, water buses, and urban trains) in the formal sector. The informal sector offers the same types of transportation (road, water, and rail), but the operators are neither structured or regulated.

The main goal is to cut down on the amount of time users spend in the transportation network, which includes time spent in vehicles, time spent waiting at stops, and time spent switching between modes and sectors.

IV. Mathematical Formulation

This section introduces the mixed-integer linear programming (MILP) model for optimizing multimodal urban transportation networks. The model minimizes total user trip time subject to operational constraints of both formal and informal sectors. Model assumptions, parameters, and decision variables are defined first, followed by the objective function and constraints.

A. Model Assumptions

The following assumptions reflect multimodal public transportation in an urban setting:

- Assumption 1, Multimodality: Users may freely combine formal and informal transport modes.
- Assumption 2, Single choice: At each stop, a user can only select one mode and one route.
- Assumption 3, Imposed waiting: If a user arrives before the next mode's scheduled departure time, they must wait until that time.
- Assumption 4, Single change: At a given time t, only one mode change is allowed per stop.
- Assumption 5, Flexibility: The total number of mode changes depends on the selected route.
- Assumption 6, Unidirectional path: The journey follows a direct path without loops.
- Assumption 7, Infrastructure: Formal modes share designated stops; informal modes operate independently without set stops.

- Assumption 8, Uniform walking speed: All users walk at a constant speed. For access time calculations, all users walk at the same speed (for example, 5 km/h).
- Assumption 9, Cross-sectoral servitude: There is a minimum required distance between formal and informal stops, which varies according to factors such as stop size, capacity, geographic location, and other urban considerations. The model accounts for these differences without explicitly identifying their causes.
- Assumption 10, Transfer points: Connections between sectors happen at the boundaries of easement areas, which means that people have to walk between them.

B. Terminology

In this section, we clarify some key terms essential for the comprehension of the model:

- Stop: A designated point where public transportation vehicles stop so that people can get on or off.
- Mode: A specific type of public transportation that a passenger uses, like a bus, a train, or a water bus.
- Segment: A part of the network that goes from one stop to the next.
- Trip: A continuous journey that a user takes without stopping, using the transportation mode. It can have one or more segments.
- Itinerary: A complete sequence of trips allowing a user to travel from an origin to a destination, possibly using more than one mode of transportation.
- Formal sector: The regulated transportation sector where all operators are officially registered enterprises that are subject to public regulation. Scheduled buses, light rail (metro), and water shuttles are some examples.
- Informal sector: The part of the transportation industry that isn't regulated and where people or unregistered businesses undertake commercial transportation services. Gbaka, Wôrô wôrô, Warren, Trotro, and Danfo are all examples from Africa.

C. Model Parameters

In Table II, we introduce the main parameters of the model.

D. Decision Variables

The following binary decision variables, in Table III, are used to formulate the constraints and objective function.

All decision variables are binary: $X_{ij}^k, Y_i^{k,k'}, W_i^{k,k'} \in \{0,1\}$ for all admissible indices. The variables X_{ij}^k select at most one mode per arc; $Y_i^{k,k'}$ indicate a mode change (and associated waiting) at stop i; and $W_i^{k,k'}$ record a cross-sector transfer when modes k and k' belong to different sectors. Constraint (11), Constraint (12) to Constraint (13), and Constraint (14) couple these variables so that waiting and sector-transfer times are activated only when consecutive arcs use different modes (and, for W, different sectors).

TABLE II. MAIN MODEL PARAMETERS

Symbol	Description
N	Set of stops.
K	Set of modes of transportation.
F	All transport sectors.
o, d	Point of departure and point of arrival.
$Change_{Max}$	Maximum number of mode changes allowed.
$t_{i,j}^k$	Time spent in the vehicle on the journey from stop i to stop j using mode k.
$u_i^{k,k'}$	Waiting time for mode change from mode k to mode k' at stop i.
$\gamma_i^{k,k'}$	Walking time required to transfer from mode k to mode k', and between sectors, at stop i.
$Z^{k,\bar{f}}$	Parameter equal to 1 if mode k belongs to the informal sector.
$Z^{k,f}$	Parameter equal to 1 if mode k belongs to the formal sector.

TABLE III. DECISION VARIABLES

Variable	Description
X_{ij}^k	Binary variable equal to 1 if the edge segment (i,j) is used with transport mode k, and 0 otherwise.
$Y_i^{k,k'}$	Binary variable equal to 1 if the user at stop i has changed mode from k to k', and 0 otherwise.
$W_i^{k,k'}$	Binary variable equal to 1 if there is a change of mode and a change of sector from k to k' at stop i, and 0 otherwise.

V. FORMULATION OF THE MODEL

A. Objective Function

The objective is to minimize the total travel time for users, which comprises three components [see Eq. (1)]:

$$\min Z = Z_1 + Z_2 + Z_3 \tag{1}$$

where, Z_1 represents the total time spent in vehicles, Z_2 represents the total waiting time at stops for connections, and Z_3 represents the total transfer time between sectors, with Eq. (2) and Eq. (3):

$$Z_1 = \sum_{i\in N}\sum_{j\in N}\sum_{k\in K} X_{i,j}^k \cdot t_{i,j}^k \tag{2}$$

$$Z_2 = \sum_{i\in N}\sum_{k\in K}\sum_{k'\in K} Y_i^{k,k'} \cdot u_i^{k,k'} \tag{3}$$

Taking the informal sector into account introduces possible changes between sectors (formal $\leftrightarrow$ informal). These transitions, constrained by formal stops, require walking between stops in different sectors. This additional time is modeled by Eq. (4):

$$Z_3 = \sum_{i\in N}\sum_{k\in K}\sum_{k'\in K} W_i^{k,k'} \cdot \gamma_i^{k,k'} \tag{4}$$

where, $\gamma_i^{k,k'}$ represents the transfer time between modes k and k' from different sectors at stop i.

B. Constraints

The constraints of the model are represented as follows:

- Single mode transportation by segment.

$$\sum_{k \in K} X_{i,j}^k \leq 1 \qquad \forall i,j \in N \times N \tag{5}$$

Constraint (5) limits the user to selecting only one mode of transport $k \in K$ for each segment $[i,j]$ travelled, from among several modes available on that section.

- Changing modes

$$\sum_{k \in K} X_{i,j}^k + \sum_{k \in K} X_{j,h}^k \leq 2 \qquad \forall i,j,h \in N \times N \times N \tag{6}$$

Eq. (6) expresses the possibility for the user to arrive at a stop j using a mode of transport k, then select another mode of transport K' to continue the journey. Typically, the mode used on segment $[i,j]$ may differ from that used on segment $[j,h]$. The upper bound of 2 of the constraint ensures that at most two modes are involved in a user's arrival and departure at a stop.

- Flow conservation

Let o be the departure stop user's and d the destination stop of the user:

$$\sum_{j \in N} \sum_{k \in K} X_{o,j}^k = 1 \tag{7}$$

$$\sum_{i \in N} \sum_{k \in K} X_{i,j}^k = \sum_{i \in N} \sum_{k \in K} X_{j,i}^k \qquad \forall j \in N \tag{8}$$

$$\sum_{j \in N} \sum_{k \in K} X_{j,d}^k = 1 \tag{9}$$

Eq. (7) to Eq. (9) enforce continuity, contiguity, and completeness of the user's journey through classic flow conservation:

- A unit flow departs from the origin o [Eq. (7)], ensuring entry into the network.
- Incoming and outgoing flows balance at all intermediate nodes [Eq. (8)], ensuring progress through the network.
- A unit flow arrives at the destination d [Eq. (9)], ensuring exit from the network.
- Capacity constraints

$$\sum_{i \in N} \sum_{k \in K} \sum_{k' \in K} Y_i^{k,k'} \leq Change_{Max} \tag{10}$$

The capacity constraint in Eq. (10) limits the number of possible mode changes in the model. This number is represented by the system parameter $Change_{\max}$. This parameter can be adjusted as needed depending on the size of the network.

- Constraint linking X and Y variables.

$$\sum_{i \in N} X_{i,j}^k + \sum_{h \in N} X_{j,h}^{k'} \leq 1 + Y_j^{k,k'} \qquad \forall j \in N,\ \forall k,k' \in K \times K,\ k \neq k' \tag{11}$$

Constraint (11) links the X variables (travel on arcs) to the Y variables (waiting at stops). A wait at stop j is possible only when arrival and departure segments at j use different modes.

- Cross-sector transfer constraints

$$W_j^{k,k'} \geq Y_j^{k,k'} + Z^{k,\bar{f}} + Z^{k',f} - 2 \qquad \forall j \in N,\ \forall k,k' \in K \times K \tag{12}$$

$$W_j^{k,k'} \geq Y_j^{k,k'} + Z^{k,f} + Z^{k',\bar{f}} - 2 \qquad \forall j \in N,\ \forall k,k' \in K \times K \tag{13}$$

These equations [Constraint (12) and Constraint (13)] model the changes from formal to informal and informal to formal, respectively.

- Constraint linking W and Y variables.

$$\sum_{k \in K} \sum_{k' \in K} W_j^{k,k'} \leq Y_j^{k,k'} \qquad \forall j \in N,\ \forall k,k' \in K \times K \tag{14}$$

Constraint (14) links changes to the corresponding waiting times.

VI. Model analysis

The suggested model offers a robust method for optimizing multimodal urban transport systems, particularly in developing nations where formal and informal transport coexist. It is founded on a mixed-integer linear programming (MILP) concept, presenting a systematic and rigorous approach to tackling the issues associated with multimodal transportation. The key components of the model are delineated below.

A. MILP Formulation and Flow Constraints

The model's core is its MILP formulation, which is especially adept at addressing the complexity and combinatorial aspects of multimodal transport optimization. The model guarantees continuity within the transportation network through flow conservation constraints. These constraints streamline the mathematical framework and enhance computational efficiency by narrowing the search space to feasible paths.

- Flow conservation: The model enforces flow conservation at the origin, intermediate stops, and destination. This guarantees that the user enters, traverses, and exits the network consistently, mirroring actual trip structure.
- Simplification through structure: The implementation of flow constraints streamlines the problem formulation, rendering it both computationally feasible and conceptually clear. This simplicity facilitates expansion while preserving the necessary flexibility to manage intricate multimodal scenarios.

B. Applicability to Multimodal Transport Systems

The model's primary strength lies in its broad application to various multimodal transport networks. The model, while developed with data from Abidjan, Côte d'Ivoire, is fundamentally adaptable to many metropolitan environments and transportation systems. The primary characteristics are as follows:

- Flexibility for different modes of transport: The model accommodates several transport modes, encompassing both formal and informal sectors, including road and lagoon possibilities. This adaptability guarantees its pertinence across diverse urban landscapes.
- By restricting mode alterations and enhancing transitions, the model can accommodate medium to large networks without considerable degradation in computing performance.
- Structured and formal design: The emphasis on formal and informal transportation modalities facilitates the establishment of a clear and dependable optimization framework. This methodology addresses the demand for systematic solutions in emerging urban areas.

C. Simplicity and Practical Applicability

The model achieves an equilibrium between mathematical precision and practical relevance. The straightforward design, enabled by clearly delineated parameters, variables, and constraints, promotes implementation. The incorporation of waiting times, mode-change limits, and operational constraints increases applicability in practical situations.

- User-centered objective: The emphasis on reducing overall journey time demonstrates a user-centered approach, consistent with the core aim of enhancing public transportation efficiency and user experience.
- Operational feasibility: The incorporation of pragmatic assumptions, including mode change constraints and predetermined stops, guarantees that the model accurately reflects the operational attributes of both formal and informal transportation systems.

D. Potential for Expansion

The model, while currently tailored for transportation systems in developing cities, possesses a modular design that offers a robust basis for future improvements. For instance:

- Incorporation of transportation expenses: Subsequent iterations may enhance the model by integrating both formal and informal transportation cost alternatives, so broadening its scope and relevance to hybrid systems.
- Advanced optimization strategies, including cut generation and boundary tightening, could enhance computational performance, enabling the model to manage larger networks.

In summary, the suggested model, utilizing the MILP approach, offers a flexible, scalable, and computationally efficient framework for optimizing multimodal transportation systems. Utilizing flow constraints, the model streamlines the problem, while preserving its adaptability and relevance across diverse situations. The simplicity, coupled with its robust mathematical base, renders it an invaluable instrument for urban transportation planning, especially in developing cities aiming to enhance the efficiency of their formal transportation systems.

VII. Computational Experiments and Data

A. Experimental Setup

The integrated MILP was implemented in Python and solved with IBM ILOG CPLEX Optimizer version 22.11 [27]. All computational experiments reported in this study use CPLEX on a workstation with an Intel Core i7-8650U CPU (2.11 GHz) and 16 GB of RAM. The model's multimodal characteristics are evidenced by the incorporation of diverse road transport modes and lagoon transport services.

B. Data Sources and Characteristics

Calibration of the computational experiments combined published mobility information for Abidjan [3], [28] with detailed scrutiny of how operators providing public transport services in the city conduct scheduled operations in practice. Access to non-public operational material was facilitated through a confidentiality agreement with the Information and Applications Research Team (ERIA). Observations drawn from that context (e.g., service structure, stop spacing, segment speeds over a multi-month horizon in 2023) were used solely to estimate distributional ranges for synthetic instance generation; no operator-identifying dataset is reproduced here, and the MILP is not solved on a published operational network graph.

The calibration for formal-sector operators providing public transport services reflects five mode types, heterogeneous stop spacing across lines, and urban link speeds up to 60 km/h, as inferred from those observations. Empirical ranges and moments were derived for segment travel times, inter-stop distances, and connection waiting. Informal-sector parameters (flexible stops, sector-specific transfer penalties $\gamma_i^{k,k'}$) were set from the same published Abidjan mobility sources and expert

judgment, because comparable operational traces for informal operators were not available under the same ERIA-mediated access regime.

C. Instance Generation

An instance generator based on Python was created to produce test instances that emulate real-world operating characteristics. The generator yields the subsequent parameters:

- `Time_mode`: Duration of in-vehicle travel for a user utilizing a certain mode of transport (minutes)
- `Time_wait`: Duration of waiting at stops for connections (minutes)
- `Number_mode`: Count of available transport modes
- `Number_change`: The maximum permissible mode changes within the network.
- `Number_stop`: The total number of stops permitted for the journey.

The generator draws parameters from uniform or truncated-normal distributions whose support was calibrated to empirically informed statistics (minimum, maximum, mean, and standard deviation of segment times and stop spacing). For each synthetic instance, arc travel times, waiting times, and the number of active arcs are sampled independently subject to network size and $Change_{Max}$. Representative calibrated ranges are:

- Travel time (`Time_mode`): uniform on $[2, 15]$ minutes (consistent with in-vehicle segment durations observed among operators providing public transport services in Abidjan)
- Waiting time (`Time_wait`): uniform on $[0, 30]$ minutes at connection stops
- Number of stops: instances with 5, 10, 15, 20, or 24 stops
- Maximum mode changes ($Change_{Max}$): 5 or 10 per instance family
- Number of modes: fixed at five (formal and informal variants per infrastructure type)

Each row in Table IV corresponds to ten random instances per ($Change_{Max}$, stop count) combination, for a total of 100 solved instances.

Table IV summarizes the characteristics of the instances. The columns denote: (1) #Stops: the number of stops per instance, (2) #Var. Total: total number of decision variables, (3) #Var. Bin.: number of binary variables, (4) #Cont: total number of constraints, (5) #Arcs: number of arcs within the interval [min, max], (6) #Inst. (ch max=5): number of instances with a maximum of 5 mode changes, and (7) #Inst. (ch max=10): number of instances with a maximum of 10 mode changes.

The dataset illustrates the model's scalability across various problem sizes. The complexity of the instance increases polynomially with the number of stops, as seen by the rising quantity of variables and constraints. The existence of cases with two distinct maximum change values (5 and 10) facilitates comparative performance evaluation under differing flexibility limitations.

D. Data Validation

Model validation was performed via a rigorous methodology. A meticulously selected test dataset was produced utilizing statistical techniques that guarantee distributions reflective of real-world scenarios. The preprocessing stages encompassed data normalization and cleansing to guarantee data quality and pertinence. Established performance criteria were utilized to assess the model on the test set, contrasting the results achieved with established optimal values when applicable.

The validation method confirmed the model's robustness and laid the groundwork for subsequent study. The validated model was then exercised on synthetic Abidjan-style network instances, producing findings that illustrate practical applicability for corridor-level urban transportation planning.

VIII. Computational Results

This section analyzes computational results from solving the proposed MILP model, focusing on the relationship between processing time and problem size (number of stops, active arcs, maximum mode transitions, and proportion of forbidden arcs).

A. Effect of Model Complexity

The duration of processing demonstrates a significant positive link with the complexity of the challenge. Instances with a greater number of stops, active arcs, and elevated maximum mode change limitations necessitate extended computation durations. The quantity of binary variables and constraints directly influences computing effort, as the solver is required to navigate an exponentially bigger solution space. This relationship is anticipated in mixed-integer programming, where the magnitude of the problem substantially affects solution time.

B. Influence of Network Size and Mode Change Limits

Fig. 1 depicts the correlation between processing duration and network attributes. The mean calculation duration increases with the quantity of stops for each value of $Change_{Max}$. This tendency is more evident when $Change_{Max} = 10$, suggesting that the increased flexibility in mode changes significantly broadens the solution space compared to $Change_{Max} = 5$.

With $Change_{Max} = 5$, the augmentation in processing time relative to the network size is less pronounced, indicating that restricting the mode changes has computational advantages, while possibly limiting solution quality. Scenarios with an increased number of stops invariably necessitate extended processing durations, irrespective of the maximum mode change limit. These observations underscore the necessity of judiciously choosing $Change_{Max}$ in accordance with problem-specific attributes and computing resource limitations.

TABLE IV. INSTANCE CHARACTERISTICS FOR FORMAL AND INFORMAL TRANSPORT SECTORS

#Stops	#Var. Total	#Var. Bin.	#Cont	#Arcs [min, max]	#Inst. (ch max=5)	#Inst. (ch max=10)
5	375	375	689	[60, 90]	10	10
10	1,000	1,000	2,224	[269, 405]	10	10
15	1,875	1,875	5,359	[637, 947]	10	10
20	3,000	3,000	10,844	[1,137, 1,709]	10	10
24	4,080	4,080	17,428	[1,663, 2,487]	10	10

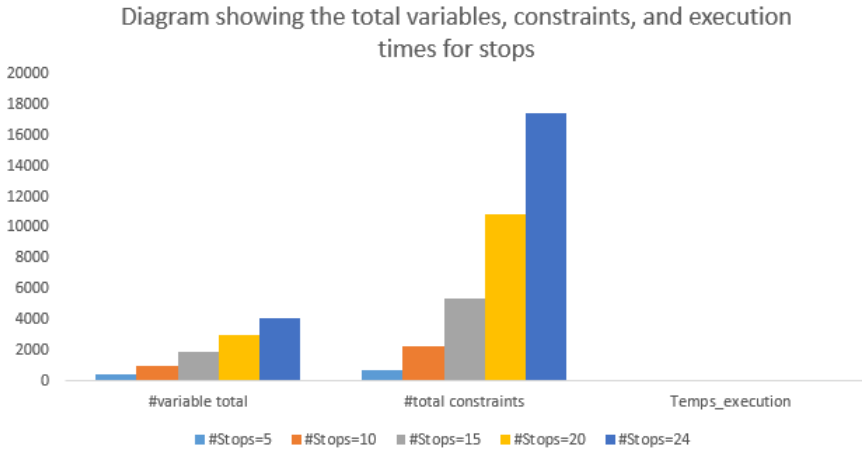


Fig. 1. Execution time and active arcs versus number of stops.

TABLE V. CHARACTERISTICS OF GENERATED INSTANCES AND CORRESPONDING CPLEX SOLVE TIMES (22.11, SAME HARDWARE AS SECTION VII-A)

Instance Family	#Stops	#Modes	#Change_max	%Forbidden Arcs	#Arcs Active	Process Time (seconds)
1	5	5	5	10	90	0.483
2	5	5	10	15	85	0.515
3	10	5	5	69	381	0.918
4	10	5	10	45	405	0.871
5	15	5	5	103	947	1.325
6	15	5	10	158	892	1.437
7	20	5	5	191	1,709	2.409
8	20	5	10	286	1,614	2.681
9	24	5	5	273	2,487	4.449
10	24	5	10	414	2,346	4.068

C. Processing Time Analysis

Table V summarizes computational results for instances of different complexity levels. The data shows clear relationships between the problem characteristics and the processing time.

The following key trends emerge from the analysis.

- Network size effect: Processing time increases with the number of stops and the number of active arcs. Larger instances with more arcs require greater computational resources, with processing time scaling approximately linearly with network size (see Figure 1).
- Mode change flexibility: Increased maximum permitted mode changes ($Change_{Max}$) broaden the solution space, leading to longer computation times. The effect is more pronounced for larger networks.
- Forbidden arcs: An increased proportion of prohibited arcs diminishes the feasible solution space while augmenting constraint complexity, resulting in variable processing times contingent upon the particular arc arrangement.

Fig. 1 illustrates that the processing time increases almost linearly with the number of stops and active arcs, signifying favorable scalability for medium-sized issue situations. This linear relationship suggests the model remains computationally tractable for practical urban transportation networks.

D. Comparative Performance Analysis

Fig. 2 compares execution times across different model configurations, illustrating how various parameter combinations affect computational performance. The analysis reveals that parameter interactions significantly influence solution time, with certain combinations leading to more efficient problem-solving.

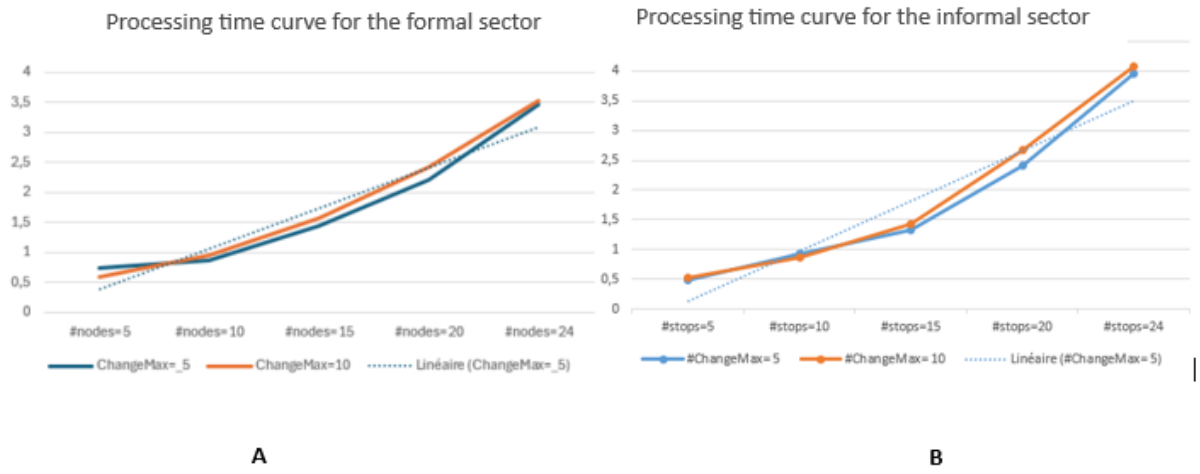


Fig. 2. Comparison of execution times across different model configurations.

IX. Sensitivity Analysis

This section reports sensitivity analysis on the impact of key parameters on MILP solution time, including $Change_{Max}$ (maximum mode changes), the number of modes, the number of stops, the percentage of forbidden arcs, and the number of active arcs.

A. Impact of Maximum Mode Changes

Increased $Change_{Max}$ significantly impacts processing time by broadening the feasible solution space. This effect is observed in instances with 5, 10, 15, and 20 stops. In small instances (5 stops), increasing $Change_{Max}$ has minimal impact on processing time, as the solution space remains tractable. For larger instances (15-24 stops), higher $Change_{Max}$ values resulted in significant increases in computation time, with processing time nearly doubling when $Change_{Max}$ grows from 5 to 10 for 24-stop instances.

This observation indicates that although increased mode changes flexibility improves solution quality, it incurs a computational cost that rises exponentially with network size. Practitioners must balance solution quality demands with available computational resources while setting the value of $Change_{Max}$.

B. Impact of Number of Transport Modes

The number of available transport modes directly influences model complexity by increasing the number of feasible options at each stop. Table V summarizes results for instances with 5 modes. Empirical analysis reveals that doubling the number of modes increases processing time by approximately 30-50% for medium-sized instances (10-15 stops). This moderate increase suggests that the model scales reasonably well with mode diversity, making it suitable for networks with multiple transport options.

C. Effect of Forbidden Arcs

The percentage of forbidden arcs introduces additional constraints that reduce the number of feasible routes while increasing computational complexity. The processing time exhibits variable behavior depending on the forbidden arc percentage. When the percentage of forbidden arcs exceeds 70%, the processing time increases slightly as the solver must navigate a more constrained solution space. However, moderate percentages (30-50%) can actually reduce processing time by eliminating infeasible paths early in the search process.

D. Travel Time Performance: Formal vs. Integrated Systems

Table VI presents a comparative analysis of travel time performance between formal-only and integrated (formal + informal) transport systems. The results demonstrate that combining formal and informal transport systems provides a more effective strategy for meeting urban mobility needs.

The analysis reveals significant structural limitations of formal-only transport systems in large African cities. While formal systems benefit from their organization, planning, and regulatory frameworks, they often fail to adequately serve all urban mobility needs, especially in peripheral regions or places with significant informal transit prevalence. The inflexibility of routes and sporadic service on specific lines diminish their efficacy as independent systems.

In contrast, informal transit provides flexibility, adaptability, and extensive access to underserved metropolitan regions, delivering responsive, locally tailored services. Nonetheless, in the absence of regulation, this sector experiences dependability challenges, safety issues, and insufficient alignment with overarching urban transportation objectives.

The intentional amalgamation of formal and informal transport elements into a cohesive multimodal framework mitigates these constraints. Computational results show that integrated systems reduce total journey time by up to 53% relative

TABLE VI. TRAVEL TIME COMPARISON: FORMAL-ONLY VERSUS INTEGRATED SYSTEMS

Instance Type	#Changes Max	#Stops	Objective Value	#Constraints	Reduction (%)
Formal only	5	5	23	308	–
Integrated	5	5	21	689	8.7
Formal only	5	10	16	1,463	–
Integrated	5	10	10	2,224	37.5
Formal only	5	15	16	4,218	–
Integrated	5	15	12	5,359	25.0
Formal only	5	20	19	9,323	–
Integrated	5	20	10	10,844	47.4
Formal only	5	24	19	15,603	–
Integrated	5	24	9	17,428	52.6

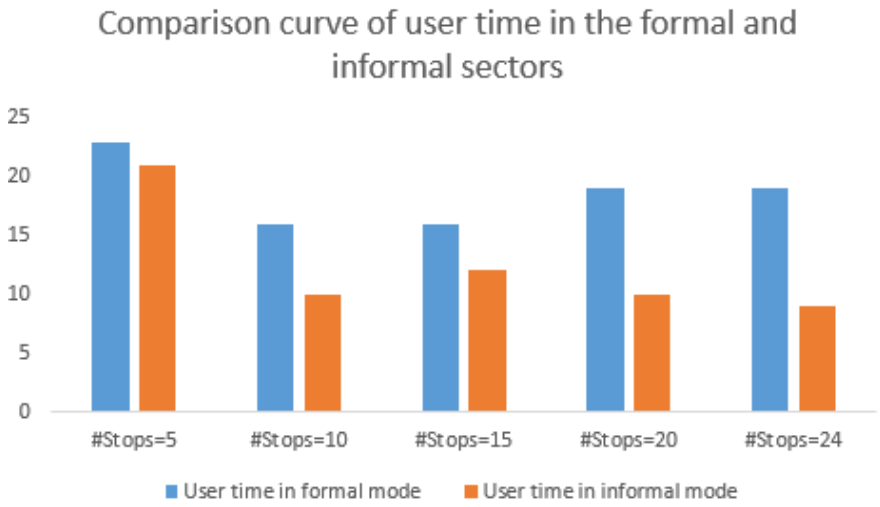


Fig. 3. Comparison of user travel time: Formal-only versus integrated (formal + informal) systems.

to formal-only configurations on the largest tested instances (24 stops; objective values 19 versus 9). Improvements are instance-dependent: for 15-stop cases the reduction is 25% (16 to 12), and Table VI reports gains between 25% and 52.6% across the tested stop counts. The headline figure of 53% therefore denotes the maximum observed improvement, not an average over all instances.

This performance enhancement arises from: 1) augmented network coverage through informal sector integration, 2) diminished waiting periods by adaptive scheduling, 3) improved accessibility to remote regions, and 4) heightened system resilience to demand variability. These findings endorse policy proposals for acknowledging and systematically incorporating informal transit into modern urban mobility planning frameworks.

Fig. 3 clearly demonstrates the performance superiority of integrated systems across various network sizes. The findings indicate steady enhancements in transit time reduction, with the advantage amplifying for more extensive networks. This study advances mathematical modeling and optimization of multimodal transport systems, while also informing broader policy debates on inclusive governance and hybrid management of transport systems in fast growing African cities.

X. COMPARATIVE ANALYSIS WITH RELATED STUDIES

This section situates the proposed MILP model relative to similar studies in the literature. The comparative analysis focuses on four key criteria: 1) model type and formulation approach, 2) optimization objective, 3) computational performance (processing time), and 4) case study complexity and scale.

A. Comparative Analysis

Table VII presents a detailed comparison with related studies. The proposed model exhibits several strengths relative to prior work.

- Travel time reduction: The integrated approach reduces total travel time by up to 53% on the largest tested instances (24 stops, 2,487 active arcs) compared with formal-only networks, and by 25–52.6% on smaller instances (Table VI). Edi et al. [26] address a related formal-sector MILP at comparable network size but without informal integration.
- Multimodal adaptability: Unlike Behiri et al. [24], who focus exclusively on rail transport, the present formulation integrates multiple transport modes (road, water, rail) from both formal and informal sectors.

TABLE VII. PERFORMANCE COMPARISON WITH RELATED STUDIES

Reference	Model Type	Optimization Criterion	Average Computation Time	Case Study Size
Proposed model	MILP with flow constraints	Total travel time	1.63 s (24 stops, 2,487 arcs)	Up to 24 stops
Edi et al. (2025) [26]	MILP with flow constraints	Total travel time	N/A (24 stops, 2,487 arcs)	Up to 24 stops
Mnif and Bouamama (2017) [25]	Multi-objective MILP	Cost and operational efficiency	3.47 s	Up to 15 stops
Behiri et al. (2016) [24]	MILP for urban rail transport	Minimized waiting time	2.3 s (10 stations, 150 arcs)	10 stations

Note. CPU times for the proposed model were obtained with CPLEX 22.11 on the hardware described in Section VII-A. Times reported for other studies used different solvers and platforms; cross-paper comparisons should be interpreted with caution.

- Computational efficiency: CPLEX solve times remain below 2 s for corridor instances up to 24 logical stops (maximum 1.63 s on the calibrated testbed), while handling explicit formal–informal sector coupling.
- Problem scale: Experiments cover corridor-level networks with up to 24 stops and 2,487 active arcs; city-wide networks would require further algorithmic development (see Section XI).

However, certain limitations and opportunities for future enhancement are identified:

- Multi-objective optimization: Unlike Mnif and Bouamama [25], our current formulation employs a single-objective function (travel time minimization). Future work could extend the model to multi-objective optimization, simultaneously considering travel time, cost, and environmental impact.
- Real-time data integration: Compared to Gonzalez et al. [16], the present formulation relies on static parameters without real-time data integration. Incorporating dynamic demand and real-time traffic conditions would enhance practical applicability.

B. Key Findings and Insights

The comparative analysis indicates that the proposed formulation offers a balanced combination of modeling transparency, computational tractability on tested instances, and relevance to developing cities where formal and informal sectors coexist.

The analysis of processing time determinants yields the following insights:

- The percentage of forbidden arcs significantly impacts processing time, as additional constraints reduce feasible solutions and increase computational effort. This relationship is consistent with theoretical expectations in constraint programming.
- While individual parameters may not show strong individual correlations, their collective interaction substantially influences computational performance. This highlights the importance of parameter optimization for efficient computation.

TABLE VIII. EXTENDED SCALABILITY EXPERIMENT (CPLEX 22.11): INTEGRATED MILP ON LARGE CALIBRATED INSTANCES ($n = 25$–100).

n	$\lvert V\rvert$	#Var.	#Constr.	Time (s)	Status
25	50	5,345	24,394	0.31	OPTIMAL
40	80	11,722	28,219	0.11	OPTIMAL
55	110	20,081	32,901	0.16	OPTIMAL
70	140	30,158	38,252	0.20	OPTIMAL
85	170	48,991	47,114	0.47	OPTIMAL
100	200	47,188	51,821	0.40	OPTIMAL

- Processing time grows with network size (Fig. 1) but remains manageable on the tested hardware for instances up to 24 stops; supplementary tests up to $n = 100$ are reported in Section XI.

XI. LIMITATIONS AND SCALABILITY

All computational results in this study were obtained with IBM CPLEX 22.11 on the workstation described in Section VII-A. The main study (Section VII to Section X) evaluates synthetic instances with 5 to 24 logical stops and up to 2,487 active arcs; the maximum observed solve time on a 1,000-instance calibrated testbed was 1.63 s. This scale is representative of sub-network or corridor planning in Abidjan but does not, by itself, demonstrate behavior on substantially larger networks. To address scalability beyond the 24-stop cap, an extended set of 100 instances was generated with the same calibration rules applied to operators providing public transport services in Abidjan, with logical stop parameters $n \in [25, 100]$ and network sizes $|V| = 2n$ (50 to 200 nodes).

Table VIII reports representative cases; all 100 runs returned an `OPTIMAL` status. Solution times increased with $|V|$ and model size but remained below 0.69 s on the largest instance ($n = 100$, $|V| = 200$, 47,188 binary variables, 51,821 constraints). Fig. 4 shows that median CPU time grows gradually with n and $|V|$ in this range, without the exponential blow-up feared for city-wide deployment at corridor scale. These results indicate that the present formulation remains tractable for medium-to-large sub-networks under static demand, while full metropolitan systems (hundreds of stops, dynamic demand) would still require decomposition, matheuristics, or rolling-horizon schemes [27].

The formulation relies on static origin–destination demand, deterministic arc travel times, and parameters fixed for the

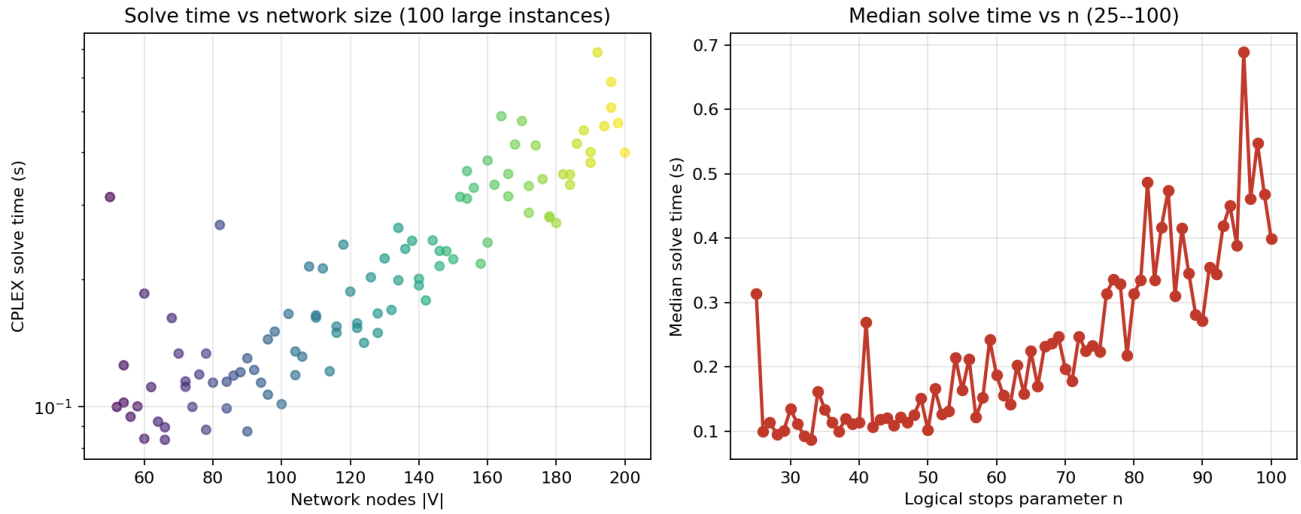


Fig. 4. CPLEX solve time versus logical stop parameter n and network size $|V|$ on 100 large instances ($n = 25$–100).

planning horizon. Extensions incorporating stochastic demand, time-dependent travel times, and real-time traffic or dispatch data would increase modeling fidelity but require additional variables, constraints, or rolling-horizon solution schemes beyond the present static MILP.

XII. Conclusion

This study presents a mixed-integer linear programming (MILP) model for optimizing multimodal urban transportation networks that explicitly accounts for both formal and informal transport sectors. The model was calibrated to operational patterns of operators providing public transport services in Abidjan and evaluated on synthetic instances, demonstrating practical relevance in developing city settings.

A. Main Contributions

The primary contributions of this research are threefold:

- Unified modeling framework: A MILP formulation integrates formal and informal transport sectors within a single optimization model, addressing a gap in the existing literature.
- Performance improvement: On synthetic instances calibrated from operators providing public transport services in Abidjan, integrated systems reduce total travel time by up to 53% relative to formal-only configurations (maximum on 24-stop cases; lower gains on smaller networks).
- Computational efficiency: CPLEX solves corridor instances ($n \leq 24$) in at most 1.63 s and an extended set with $n \in [25, 100]$ and $|V| \leq 200$ in at most 0.69 s (Section XI), while metropolitan-scale dynamic planning remains future work.

B. Key Findings

Sensitivity analysis reveals that processing time is primarily influenced by: 1) the number of stops, 2) the maximum number of mode changes allowed ($Change_{Max}$), and 3) the number of active arcs. Higher values of $Change_{Max}$ and number of modes increase computational complexity by expanding the solution space, while higher percentages of forbidden arcs tighten constraints, adding to computational complexity. Despite these challenges, the model remains solvable within seconds on the tested instance sizes, with processing time increasing with network size as expected for MILP (Section XI).

C. Future Research Directions

Several promising directions for future research are identified:

- Advanced solution techniques: Employing commercial solvers such as CPLEX [27] or Gurobi, along with specialized techniques such as cut generation and bound tightening, could improve computational efficiency for larger problem instances.
- Multi-objective optimization: Extending the model to simultaneously optimize travel time, transportation costs, and environmental impact would enhance practical relevance.
- Dynamic and stochastic elements: Incorporating real-time demand fluctuations, stochastic travel times, and dynamic network conditions would increase model applicability to operational planning.
- Cost integration: Incorporating transportation cost parameters for different routes and modes would enable comprehensive cost-benefit analysis and support policy decision-making.
- Network design extensions: Expanding the model to include network design decisions, such as stop location and route planning, would provide a more comprehensive optimization framework.

The proposed model provides a solid foundation for optimizing multimodal urban transportation systems in developing cities, where formal and informal sectors coexist. The

demonstrated performance improvements and computational tractability make it a valuable tool for urban transportation planners and policymakers seeking to enhance mobility while recognizing the complementary roles of both transport sectors.

REFERENCES

[1] UN-Habitat, *World Cities Report 2022: Envisaging the Future of Cities*. Nairobi, Kenya: UN-Habitat, 2022, urban population growth projections; statistical annex. [Online]. Available: https://unhabitat.org/wcr/2022/

[2] R. Cervero and A. Golub, "Informal transport: A global perspective," *Transport Policy*, vol. 14, no. 6, pp. 445–457, 2007.

[3] Autorité de la Mobilité Urbaine dans le Grand Abidjan (AMUGA), "La mobilité en chiffres – parts modales en 2020," Official statistical report, Abidjan, Côte d'Ivoire, 2020, accessed 27 May 2025. [Online]. Available: https://amuga.ci/mobilite-en-chiffres/

[4] P. Sakho, "Les acteurs du système de transports et les politiques publiques de modernisation au sénégal: Les acteurs du transport informel face à la mise en œuvre du BRT de dakar," Research report, Dakar, Senegal, 2021, accessed 27 May 2025. [Online]. Available: https://www.researchgate.net/publication/356916717

[5] Lagos Metropolitan Area Transport Authority, "Lagos transport policy: Integration of informal modes," Transport policy document, Lagos, Nigeria, 2022, accessed 27 May 2025. [Online]. Available: https://www.lamata-ng.com

[6] CongoInter, "Kinshasa adopte les taxis fluviaux pour désengorger la circulation," Online news article, Mar. 2025, accessed 27 May 2025. [Online]. Available: https://congointer.info/2025/03/11/kinshasa-adopte-les-taxis-fluviaux-pour-desengorger-la-circulation/

[7] E. Ravalet and S. Vincent-Geslin, "Mobilités, activités et territoires du quotidien à niamey: La ségrégation au-delà des espaces résidentiels?" *Espace Populations Sociétés*, no. 1–2, 2015. [Online]. Available: https://journals.openedition.org/eps/6067

[8] K. Kepaptsoglou and M. Karlaftis, "Transit route network design problem," *Journal of transportation engineering*, vol. 135, no. 8, pp. 491–505, 2009.

[9] D. Bertsimas, Y. S. Ng, and J. Yan, "Data-driven transit network design at scale," *Operations Research*, vol. 69, no. 4, pp. 1118–1133, 2021.

[10] W. Fan and R. B. Machemehl, "Optimal transit route network design problem with variable transit demand: genetic algorithm approach," *Journal of transportation engineering*, vol. 132, no. 1, pp. 40–51, 2006.

[11] ——, "Using a simulated annealing algorithm to solve the transit route network design problem," *Journal of transportation engineering*, vol. 132, no. 2, pp. 122–132, 2006.

[12] ——, "Tabu search strategies for the public transportation network optimizations with variable transit demand," *Computer-Aided Civil and Infrastructure Engineering*, vol. 23, no. 7, pp. 502–520, 2008.

[13] ——, "Bi-level optimization model for public transportation network redesign problem: Accounting for equity issues," *Transportation Research Record*, vol. 2263, no. 1, pp. 151–162, 2011.

[14] M. A. B. Drummond, "Optimisation d'un système multimodal de transport de marchandises," 2014.

[15] P. Sitek and J. Wikarek, "Cost optimization of supply chain with multimodal transport," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2012, pp. 1111–1118.

[16] A. B. R. González, J. J. V. Díaz, M. R. Wilby, and R. F. Pozo, "Data-driven performance evaluation framework for multi-modal public transport systems," *Sensors*, vol. 22, no. 1, p. 17, December 2021. [Online]. Available: https://doi.org/10.3390/s22010017

[17] A. Mogbojuri, O. Olanrewaju, and A. Adeyeye, "Linear programming model for scheduling bus rapid transit in lagos state, nigeria," *Nigerian Journal of Technology*, vol. 42, no. 2, pp. 236–243, 2023.

[18] H. Ma, X. Li, and H. Yu, "Single bus line timetable optimization with big data: A case study in beijing," *Information Sciences*, vol. 536, pp. 53–66, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520302887

[19] E. Sezgin and Ş. Köfteci, "Public transportation planning using linear goal programming: A case study of antalya," *Sigma Journal of Engineering and Natural Sciences*, vol. 36, no. 2, pp. 393–405, 2018. [Online]. Available: https://dergipark.org.tr/en/pub/sigma/issue/35818/399911

[20] S. Das Gupta, J. K. Tobin, and L. Pavel, "A two-step linear programming model for energy-efficient timetables in metro railway networks," *Transportation Research Part B*, vol. 93, pp. 57–74, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261516304830

[21] E. Berhan, "Linear programming optimization techniques for Addis Ababa public bus transport," in *Optimization of Fuel Cell and Hybrid Vehicle Components*. London, United Kingdom: IntechOpen, 2020. [Online]. Available: https://doi.org/10.5772/intechopen.93629

[22] K. Huang, J. Wu, H. Sun, X. Yang, Z. Gao, and X. Feng, "Timetable synchronization optimization in a subway–bus network," *Physica A: Statistical Mechanics and its Applications*, vol. 608, p. 128273, 2022.

[23] A. Eslami, Y. Shafahi, and S. Bafandkar, "Optimizing and synchronizing timetables in an urban subway network considering trains' speed profiles and skip-stop strategy," *Journal of Rail Transport Planning & Management*, vol. 34, p. 100520, 2025.

[24] W. Behiri, O. Ozturk, and S. Belmokhtar-Berraf, "Urban freight by rail: A milp modeling for optimizing the transport of goods," in *The 6th International Conference on Information Systems, Logistics and Supply Chain*, Bordeaux, France, Jun. 2016. [Online]. Available: https://hal.science/hal-01740892

[25] M. Mnif and S. Bouamama, "A multi-objective formulation for multimodal transportation network's planning problems," in *2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. IEEE, sep 2017.

[26] Y. F. A. Edi, O. Koné, and K. H. Edi, "A new mixed-integer linear programming approach to the problem of multimodal urban transit," *Statistics, Optimization & Information Computing*, Apr. 2025. [Online]. Available: http://www.iapress.org/index.php/soic/article/view/2158

[27] IBM, "Ibm ilog cplex optimization studio user's manual," *Version*, vol. 12, no. 1987-2018, p. 1, 1987.

[28] Autorité de la Mobilité Urbaine du Grand Abidjan, "Mobilité en chiffres – temps de correspondance entre modes de transport," 2025, consulté en juin 2025. [Online]. Available: https://amuga.ci/mobilite-en-chiffres/

Artificial Intelligence in Motorsport: A Scoping Review of Applications, Challenges, and Research Gaps

Radu C. Lucaciu[1], Cornelia Aurora Győrödi[2]
Doctoral School of Engineering Studies-Department of Computer Science and Information Technology, University of Oradea, Oradea, Romania[1]
Department of Computer Science and Information Technology, University of Oradea, Oradea, Romania[2]

***Abstract*—This study presents a state-of-the-art review of Artificial Intelligence and Machine Learning applications in motorsport, with a particular focus on Formula 1. As modern racing generates increasingly large volumes of high-frequency telemetry data, traditional physics-based analysis methods have proven insufficient to capture the complex, non-linear relationships between vehicle, driver, and environment. AI has emerged as a critical tool for extracting actionable insights from this data across six core domains: vehicle performance optimization, race strategy, autonomous racing and real-time decision support, telemetry and race data analysis, driver coaching and simulation, and predictive maintenance. This review further examines transferable methodologies from adjacent fields — particularly autonomous driving research — and assesses their relevance to high-performance motorsport contexts. A systematic analysis of peer-reviewed publications and preprints reveals a highly uneven distribution of research, with race strategy and vehicle optimization relatively well-covered, while predictive maintenance and computer vision applications remain largely unexplored.**

***Keywords*—Artificial Intelligence; Machine Learning; motorsport; Formula 1; scoping review; PRISMA-ScR; race strategy optimization; telemetry analysis**

I. Introduction

Recently, motorsport competition has developed substantially under the influence of technology and data-driven decisions. The current telemetry systems and high precision sensors, as well as modern and performance data collection infrastructure provide a way to collect massive volumes of data in near-real time. Different variables, ranging from engine, suspension or gearbox parameters to tyre temperature and degradation, driver behaviour or environmental temperature and pressure are collected and used for complex analysis to provide technical and strategic insights but the volume, data-generation speeds and both linear and non-linear dependencies between variables have outgrown the capabilities of traditional methods of analysis. In this context, Artificial Intelligence has emerged as an ideal tool for high performance motorsport, capable of transforming raw data into valuable insights for performance optimisation.

To ensure conceptual clarity and consistency throughout this review, the key terms used in the literature are defined here in a hierarchical manner. Artificial Intelligence (AI) is used as an umbrella term encompassing any computational method capable of extracting complex patterns, generating predictions, or making decisions based on data, including both traditional and modern approaches. Machine Learning (ML) refers to the subset of AI in which models learn from data without being explicitly programmed, covering supervised learning (e.g., regression, classification), unsupervised learning (e.g., clustering), and reinforcement learning (RL). Deep Learning (DL) is a further subset of ML that employs multi-layered neural network architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models. Finally, the term "data-driven models" is used more broadly to include both ML-based approaches and non-ML statistical methods such as multivariate regression or principal component analysis, provided they learn relationships directly from data rather than from first-principles physics. The relationships between these terms are hierarchical: DL $\subset$ ML $\subset$ AI, while data-driven models partially overlap with ML but also include classical statistical methods. Throughout this study, the most specific applicable term is used in each context; when referring to the overall field or to studies that employ multiple approaches, the general term "AI" is used.

However, adopting AI in motorsport is not a simple process. Racing cars operate at the extreme physical limits: heavy acceleration and deceleration, massive cornering speeds and lateral G-force loads, rapid degradation of tyres and complex strategic decisions like optimal pit strategies, the correct choice of tyre compounds and even technical choices such as mechanical and aerodynamic vehicle setup. In these conditions, standard physics-based models can become too complex or expensive (e.g. CFD simulations or multi-variable analysis) and can sometimes miss non-linear correlations between various components: car, driver and environment. AI, on the other hand, is extremely versatile at detecting and extracting these patterns from vast amounts of data, offering a very pragmatic way of modeling complexity.

Formula 1 has long been considered the technological pinnacle of motorsport and it makes sense that academic research and motorsport industry have converged in multiple areas. Deep reinforcement learning has been applied to autonomous racing using telemetry data [1], sim racing telemetry has been used to predict performance with high accuracy [2], and autonomous racing projects have become real-life laboratories for testing perception, planning and control models at high speeds and variable conditions [3]. AI is no longer just a theoretical research subject: it is a mandatory tool which confers a real competitive advantage, capable of influencing various aspects of performance motorsport, from strategic choices to car setup

and even driver coaching and selection.

Despite this growing adoption, no comprehensive scoping review has systematically mapped the extent and nature of AI research across the various motorsport application domains. Existing reviews tend to focus on narrow sub-topics: Betz et al. [3] provide a thorough survey of autonomous racing but do not address race strategy, telemetry analysis, or predictive maintenance; Kuutti et al. [4] and Elallid et al. [5] survey deep learning and reinforcement learning for autonomous driving, a field adjacent to motorsport but with fundamentally different performance objectives; and Arena et al. [6] review predictive maintenance in the broader automotive sector without addressing the specific constraints of competitive racing. What differentiates the present analysis from these prior works is its cross-domain scope: rather than focusing on a single technique or application area, this review maps the full landscape of AI research in motorsport, identifies the connections between application domains, and highlights the gaps that remain unexplored. A scoping review approach, following the PRISMA-ScR framework, is particularly well-suited to this purpose given the heterogeneous body of literature available — spanning peer-reviewed journals, conference proceedings, and publicly available preprints — and the exploratory nature of the research landscape.

The primary objective of this scoping review is to systematically map the existing body of research on AI and ML applications in motorsport across defined thematic categories, providing a comprehensive overview of how these technologies are currently being applied and studied. Building on this mapping, the review further aims to identify which application domains are well-covered in the literature and which remain underexplored, to examine transferable methodologies from adjacent fields — particularly autonomous driving — and assess their relevance to high-performance motorsport contexts, and to highlight key research gaps and propose directions for future investigation.

The review is structured around the PCC (Population, Concept, Context) framework: the *Population* comprises published academic works addressing motorsport engineering, autonomous driving and competition; the *Concept* is the application of AI and ML techniques; and the *Context* is motorsport competition, with a particular focus on Formula 1 and related high-performance racing series.

II. Methods

A. Protocol and Registration

This scoping review was conducted following the methodological framework proposed by Arksey and O'Malley [7] and refined by Levac et al. [8], and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews (PRISMA-ScR) checklist [9]. No formal protocol was registered prior to conducting the review.

B. Eligibility Criteria

Sources of evidence were included if they met the following criteria: 1) addressed the application of AI or ML methods to any aspect of motorsport or high-performance vehicle operation; 2) were published in peer-reviewed journals, conference proceedings, or as verifiable preprints and theses; or 3) were relevant cross-domain works from autonomous driving research with plausible conceptual relevance to motorsport contexts.

Sources were excluded if they: 1) focused exclusively on conventional rule-based or purely physics-based approaches without any data-driven component; 2) addressed motorsport only in a marginal or illustrative capacity; or 3) were not available in English.

The primary temporal focus was literature published from 2020 onwards, reflecting the period during which deep learning and large-scale data-driven approaches became systematically adopted in both academic research and applied motorsport engineering. However, selected earlier works were included where they provided essential historical context or represented foundational contributions to the field.

C. Information Sources

Literature was retrieved from three primary academic databases: Scopus, IEEE Xplore, and Google Scholar. The searches were conducted on February 7th 2026 with a further refining search on May 23rd 2026. Reference lists of included articles were also hand-searched to identify additional relevant sources.

D. Search Strategy

Searches were conducted using combinations of the following keywords: *motorsport*, *Formula 1*, *artificial intelligence*, *Machine Learning*, *deep learning*, *reinforcement learning*, *telemetry*, *race strategy*, *tyre degradation*, *autonomous racing*, *driver coaching*, and *predictive maintenance*. Boolean operators (`AND`, `OR`) were applied to combine domain-specific and methodological terms, for example: *"Machine Learning AND motorsport"* or *"reinforcement learning AND race strategy"*. The full search string for Scopus was: `TITLE-ABS-KEY((``artificial intelligence'' OR ``machine learning'' OR ``deep learning'' OR ``reinforcement learning'') AND (``motorsport'' OR ``Formula 1'' OR ``race strategy'' OR ``autonomous racing''))`. Table I presents the search strategy per database.

E. Selection of Sources of Evidence

All records retrieved from the database searches were first screened by title and abstract for relevance to AI/ML applications in motorsport or closely related high-performance vehicle domains. Full texts of potentially relevant articles were then assessed against the eligibility criteria. To enhance consistency, the eligibility criteria were pilot-tested on a subset of 15 randomly selected records prior to full screening, and the criteria were iteratively refined based on this pilot phase. In cases of ambiguity regarding inclusion, articles were retained for full-text review to minimize the risk of excluding relevant work. Additionally, a random sample of 10 previously screened records was re-screened at the end of the process to verify consistency of the applied criteria.

TABLE I. SEARCH STRATEGY PER DATABASE

Database	Records	Date	Search query / strategy
Scopus	51	7 Feb 2026	`TITLE-ABS-KEY(("artificial intelligence" OR "machine learning" OR "deep learning" OR "reinforcement learning") AND ("motorsport" OR "Formula 1" OR "race strategy" OR "autonomous racing"))`
IEEE Xplore	38	7 Feb 2026	`("artificial intelligence" OR "machine learning" OR "deep learning" OR "reinforcement learning") AND ("motorsport" OR "Formula 1" OR "race strategy" OR "autonomous racing")` (Command Search, all metadata)
Google Scholar*	55	7 Feb 2026	Keyword combinations of `machine learning`, `deep learning` and `reinforcement learning` with `motorsport`, `Formula 1`, `race strategy` and `autonomous racing`; first 100 results per query screened; date range 2003–2026.

Google Scholar does not support field-restricted (title/abstract/keyword) searching or fully nested Boolean queries; simplified keyword combinations were, therefore, used and results screened by relevance ranking.

F. Data Charting Process

A data charting form was developed to extract relevant information from each included source. The form was iteratively refined as the reviewer became increasingly familiar with the literature. Data were charted by the primary reviewer and verified through re-reading of the source material. The charting categories included: author(s), year of publication, source type, AI/ML method(s) employed, motorsport application domain, racing context (e.g., Formula 1, Formula Student, autonomous racing), data source used (e.g., real telemetry, simulator, synthetic), and key findings. Additionally, research gaps were identified through a thematic frequency analysis of the charted data. After categorizing all included sources into the six predefined application domains, areas with notably fewer studies relative to other domains were flagged as underexplored. This quantitative threshold-based approach was complemented by a qualitative assessment of the depth of existing research within each domain, considering factors such as methodological diversity, data source variety, and the presence or absence of real-world validation.

G. Data Items

The following data items were extracted from each included source: (1) bibliographic information (authors, year, publication venue); (2) study type (journal article, conference paper, thesis, preprint); (3) AI/ML technique(s) used (e.g., deep learning, reinforcement learning, supervised learning, ensemble methods); (4) primary application domain (vehicle performance, race strategy, telemetry analysis, driver coaching, predictive maintenance, or autonomous racing); (5) motorsport context (Formula 1, Formula Student, MotoGP, autonomous racing, general motorsport); (6) data source (real-world telemetry, simulator data, synthetic/generated data, public API); (7) research approach (simulation-based, experimental, conceptual/survey, or hybrid); (8) technical maturity level, classified using a simplified scale: conceptual or theoretical, simulation-validated, experimentally tested with real data, or deployed in competition; and (9) key findings and contributions.

III. RESULTS

A. Selection of Sources of Evidence

The database searches yielded a total of 144 records (Scopus: 51, IEEE Xplore: 38, Google Scholar: 55). Eight additional sources were identified through reference-list manual search and supplementary targeted searches. After removing 26 duplicates, 126 records were screened by title and abstract, of which 49 were excluded for not meeting the eligibility criteria (primarily papers addressing general AI/ML, sports medicine, or other non-motorsport domains). The remaining 77 articles were assessed in full text, leading to the exclusion of 27 sources (12 focused exclusively on physics-based methods without a data-driven component, 10 addressed motorsport only marginally, 3 were not available in English, and 2 were duplicates of included works by the same authors). In total, 50 sources were included in this scoping review. The selection process is illustrated in the PRISMA flow diagram (Fig. 1).

B. Characteristics of Sources of Evidence

The 50 included sources spanned the period from 2003 to 2026, with the majority (n=38) published from 2020 onwards. Source types included journal articles (n=30), conference papers (n=11), theses and dissertations (n=6) and preprints (n=3). The geographic distribution of first authors was varied, with contributions from Europe, North America, Asia, and the Middle East. The motorsport contexts addressed included Formula 1, Formula Student, MotoGP, Endurance Racing, autonomous racing (e.g., Indy Autonomous Challenge, Roborace), and general motorsport engineering.

C. Data Charting Results

Table II and Table III presents the data extracted from each included source. Papers were categorized into six thematic application domains: Vehicle Performance Optimization (VP), Race Strategy and Real-Time Decisions (RS), Telemetry and Race Data Analysis (TA), Driver Coaching and Simulation (DC), Predictive Maintenance and Reliability (PM), and Autonomous Racing (AR).

D. Synthesis of Results

1) Distribution across application domains: The distribution of the 50 reviewed sources across the six thematic categories is presented in Fig. 2. The six application domains were defined a priori based on a preliminary scan of the literature and reflect the principal areas in which AI methods are applied in motorsport: Vehicle Performance Optimization (VP), Race Strategy and Real-Time Decisions (RS), Telemetry and Race Data Analysis (TA), Driver Coaching and Simulation (DC), Predictive Maintenance and Reliability (PM), and Autonomous Racing (AR). Papers with cross-cutting contributions were assigned to the single category most closely matching their

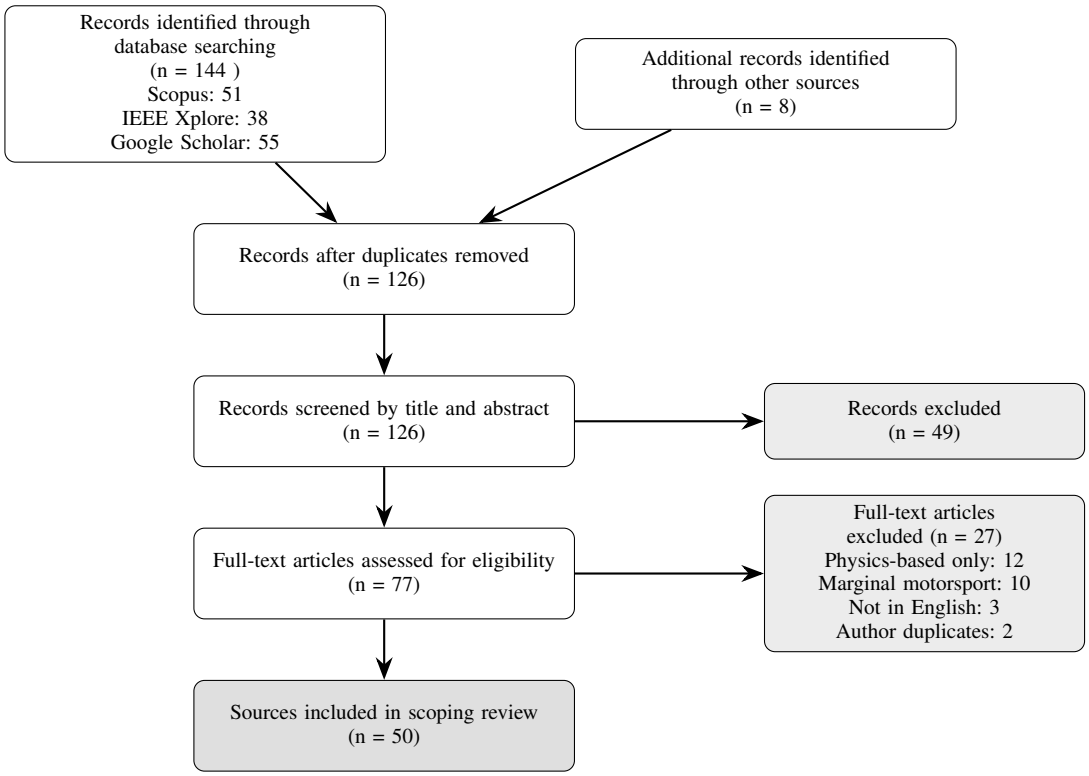


Fig. 1. PRISMA flow diagram illustrating the source selection process.

primary research focus, based on the declared objectives and the principal contribution of each study. In cases, where a paper contributed to multiple domains (e.g., a study combining vehicle dynamics modeling with race strategy), the dominant methodological contribution was used as the classification criterion. As shown in Fig. 2, Race Strategy dominates the current research landscape (n=14), followed by Vehicle Performance Optimization and Autonomous Racing (n=12 each), while Predictive Maintenance remains notably underrepresented (n=2), indicating a significant gap in the literature.

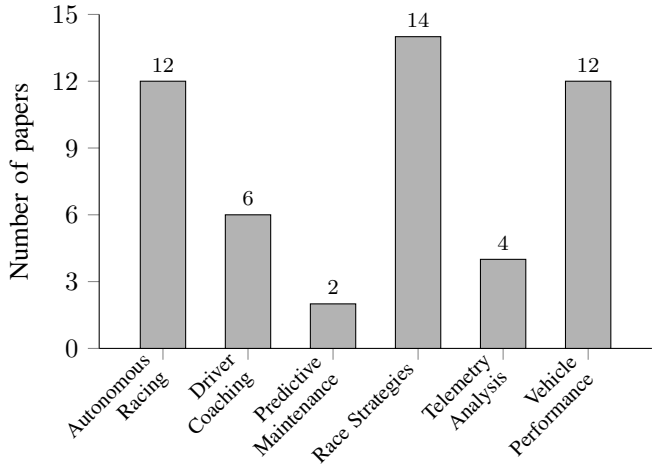


Fig. 2. Distribution of reviewed papers across application categories.

2) Research approach and technical maturity: To assess the current state of AI development in motorsport beyond thematic coverage, each included source was classified by its research approach (simulation-based, experimental, conceptual/survey, or hybrid) and by its technical maturity level, using a simplified four-level scale: conceptual or theoretical, simulation-validated, experimentally tested with real-world data, or deployed in competition. The cross-tabulation of these two dimensions is presented in Table IV.

The analysis reveals that the majority of included studies remain at the simulation-validated stage (n=23), relying on simulated or synthetic data to train and evaluate AI models. Only 14 studies have been experimentally validated using real-world data, and merely 5 have demonstrated deployment in actual competition or real-time racing environments — notably Kabzan et al. [37] (learning-based MPC for autonomous racing), Betz et al. [46] (full-stack autonomous racing at the Indy Autonomous Challenge), and Wurman et al. [31] (RL agent outperforming human drivers in Gran Turismo Sport). Eight sources are conceptual or survey papers that synthesize existing knowledge without introducing new experimental results.

This distribution indicates that while AI research in motorsport is growing rapidly, the field remains largely in the simulation and proof-of-concept phase. The gap between simulation-validated methods and real-world deployment is substantial, reflecting both the technical challenges of transferring AI models to the extreme conditions of racing vehicles and the data-access constraints imposed by the secretive nature of the sport. Future work should prioritize bridging this simulation-to-reality gap, particularly through sim-to-real transfer techniques and the

TABLE II. DATA EXTRACTION TABLE OF INCLUDED SOURCES (PRISMA-SCR CHARTING)

Author(s)	Year	Source Type	AI/ML Method(s)	Cat.	Motorsport Context	Data Source	Key Findings
Remonda et al.[1]	2021	Preprint	Deep RL (DDPG)	AR	Autonomous racing	Simulator	DDPG agent trained on telemetry data able to autonomously drive race cars
Hojaji et al.[2]	2024	Journal	ML (ensemble, feature selection)	DC	Sim racing	Simulator (ACC)	97.19% prediction accuracy for lap performance; lateral accel., steering angle most significant
Betz et al.[3]	2022	Journal	Survey (RL, MPC, CV)	AR	Autonomous racing	Multiple	Comprehensive survey of autonomous vehicle racing; perception, planning, control
Ishii et al.[10]	2010	Journal	Electronic control	VP	Formula 1	Real telemetry	Early electronic control systems for F1; foundational data acquisition
Besselink et al.[11]	2008	Conference	Data-driven vehicle dynamics	VP	General	Simulation	First integration of data-driven methods for yaw dynamics optimization
Richardson[12]	2017	Thesis	Statistical modeling	VP	General motorsport	Simulation	Tyre performance evolution modeling in real-time simulation environment
Piccinotti[13]	2019	Thesis	Open-loop optimization	RS	Formula 1	Simulation	Open-loop planning approach for F1 race strategy identification
Vesel[14]	2015	Journal	Algorithmic optimization	VP	General motorsport	Simulation	Automated racing line optimization for multiple vehicle types and circuits
Stoll et al.[15]	2013	Conference	Computer Vision	TA	General motorsport	Sparse data	Racecar tracking and visualization using sparse data techniques
Cressell & Tornberg[16]	2018	Thesis	Object recognition, SLAM	AR	Formula Student	Real-world	Object recognition and localization for autonomous racecar
Choo[17]	2015	Thesis	Predictive analytics	RS	Professional motorsport	Race data	Real-time decision-making analytics for professional race strategy
Diasinos et al.[18]	2017	Journal	Numerical/CFD + data	VP	Formula-type	Simulation	Aerodynamic analysis of ride height effects on wing-wheel interactions
Zhang & Zerihan[19]	2003	Journal	Experimental + numerical	VP	General racing	Wind tunnel	Aerodynamics of double-element wing in ground effect
Akin[20]	2024	Thesis	Data-driven optimization	VP	Formula Student	Real + simulator	Data-driven vehicle performance optimization for Formula Student
Sasikumar et al.[21]	2025	Journal	Deep learning (Bi-LSTM, TCN-GRU, CNN)	RS	Formula 1	Public API (FastF1)	Bi-LSTM best for pitstop prediction (F1=0.81); uses SMOTE for class balancing
Rao[22]	2025	Thesis	Ensemble learning	RS	Formula 1	Race data	Predictive model for pitstop strategy using ensemble methods
Rondelli[23]	2022	Journal	Neural networks	RS	Formula 1	Race data	Neural networks to predict tyre strategy in F1
Shakila & Baalaji[24]	2025	Conference	Big data analytics, predictive models	RS	Formula 1	Race data	Survey on optimizing tyre and brake performance using big data analytics
Avignone et al.[25]	2025	Journal	Data mining, visualization	TA	Racing (general)	Telemetry	Mine4Race toolset for telemetry analysis and visualization
Basile[26]	2024	Preprint	Anomaly detection	TA	MotoGP	Telemetry	Anomaly detection techniques applied to MotoGP telemetry data
Theissler et al.[27]	2021	Journal	ML (multiple methods)	PM	Automotive / motorsport	Sensor data	Predictive maintenance via ML; use cases and challenges in automotive
Arena et al.[6]	2022	Journal	Literature review (ML)	PM	Automotive / motorsport	Multiple	Comprehensive review of predictive maintenance in automotive sector
Kuutti et al.[4]	2019	Journal	DL survey (DRL, CNN, RNN)	AR	Autonomous vehicles	Multiple	Survey of DL for autonomous vehicle control; layered architecture
Di & Shi[28]	2021	Journal	RL, physics-AI hybrid	AR	Autonomous vehicles	Multiple	Mixed-autonomy control; from physics-based to AI-guided policy learning
Elallid et al.[5]	2022	Journal	DL, RL survey	AR	Autonomous driving	Multiple	Comprehensive survey of DL and RL in autonomous driving
Pérez-Gil et al.[29]	2022	Journal	DRL (DQN, DDPG)	DC	Autonomous (CARLA)	Simulator	DRL-based vehicle control comparable to classical LQR controllers
Reda et al.[30]	2023	Journal	Hybrid ML (SL + RL)	DC	Autonomous driving	Simulator	Hybrid approach balancing accuracy, finesse, and response time

Categories: VP = Vehicle Performance Optimization; RS = Race Strategy and Real-Time Decisions; TA = Telemetry and Race Data Analysis; DC = Driver Coaching and Simulation; PM = Predictive Maintenance and Reliability; AR = Autonomous Racing.

development of standardized benchmarks.

3) Brief overview of AI usage in motorsport: Although data collection in Formula 1 has been implemented starting with the mid-1980s when teams started using digital systems for data acquisition, analysis was still being done through traditional statistical methods or rule-based systems. Telemetry was collected using burst transmissions from the car during races [10]. During the 1990s, usage of digital technology picked up considerably as modern sensors became prevalent with the advent of active-ride cars [54]. Public research during this period was already treating the potential benefits of algorithmic optimization of vehicle performance.

The early 2000s marked the switch from purely physics-based analysis to hybrid models (physical–data driven) as a response to the greater complexity of racing vehicles. Even though modern Machine Learning was not widely used, the emergence of multivariate regression, non-linear optimisation and solid statistical models was evident. Besselink [11] integrated the first data-driven methodologies, allowing teams to adjust simulation strategies based on historical data. The

TABLE III. DATA EXTRACTION TABLE OF INCLUDED SOURCES — CONTINUED

Author(s)	Year	Source Type	AI/ML Method(s)	Cat.	Motorsport Context	Data Source	Key Findings
Wurman et al.[31]	2022	Journal	Deep RL (SAC)	AR	Gran Turismo (sim)	Simulator	RL agent outperformed champion human drivers in GT Sport
Fuchs et al.[32]	2020	Journal	Deep RL	AR	Gran Turismo (sim)	Simulator	Super-human performance in GT Sport via deep RL
Liu & Fotouhi[33]	2020	Journal	ANN, MCTS	RS	Formula E	Simulation	ANN + Monte Carlo tree search for Formula-E race strategy
Garlick & Bradley[34]	2021	Journal	ML (trajectory opt.)	VP	General motorsport	Simulation	Real-time optimal trajectory planning and lap time simulation
Boettinger & Klotz[35]	2023	Preprint	RL, simulation	RS	Nürburgring	Simulator	Comprehensive race simulation for AI strategy decisions
Hojaji et al.[36]	2022	Conference	ML (classification)	DC	Sim racing	Simulator	ML modeling of driver performance in simulated racing
Kabzan et al.[37]	2019	Journal	Learning-based MPC	AR	Autonomous racing	Real-world	Learning-based MPC for autonomous racing
Wurman et al.[38]	2022	Journal	RL survey	AR	Autonomous racing	Multiple	Challenges and opportunities of RL in autonomous racing
Heilmeier et al.[39]	2020	Journal	ANN (strategy)	RS	Circuit motorsport	Race data	Virtual Strategy Engineer: ANN for race strategy decisions
Boiarov et al.[40]	2023	Conference	Computer Vision, ML	TA	General racing	Race photos	RaceLens: ML-based racing photo analysis application
Veerasagar et al.[41]	2025	Journal	RL	RS	Formula 1	Race data	F1 strategy predictor using reinforcement learning
Paparusso et al.[42]	2022	Journal	Stochastic opt.	RS	Hybrid racing	Simulation	Competitors-aware stochastic lap strategy optimization
Remonda et al.[43]	2021	Journal	DRL, behavioral	DC	Professional racing	Simulator	Comparing human vs. autonomous driving in racing simulator
Kavitha et al.[44]	2025	Conference	Survey (AI in F1)	RS	Formula 1	Multiple	Overview of AI applications in Formula One racing
Raji et al.[45]	2022	Conference	Motion planning, MPC	AR	Autonomous racing	Real-world	Multi-vehicle autonomous racing at high speeds
Betz et al.[46]	2022	Journal	Full-stack autonomous	AR	Indy Auton. Challenge	Real-world	Complete autonomous racing software for IAC
Hojaji et al.[47]	2023	Conference	AI, telemetry	DC	Sim racing	Simulator	AI analysis of driving behaviour using telemetry data
Jiménez Elbal et al.[48]	2024	Journal	Optimization (ML)	VP	Electric racing	Simulation	Simultaneous vehicle design and control optimization
Bonini et al.[49]	2023	Conference	ANN, Kalman filter	VP	MotoGP	Real-world	ANN and Kalman filter estimate carbon-brake friction coefficient (μ) for accurate in-race braking-torque computation; validated with real track data
Bonini et al.[50]	2026	Journal	NARX, ANN	VP	MotoGP	Real-world	NARX and ANN estimate the disc/pad friction coefficient (μ) from telemetry; combined with a physical model to compute front-wheel braking torque; validated on multi-track real data
Jain & Morari[51]	2020	Conference	Bayesian optimization	VP	General motorsport	Simulation	Computes the racing line (path and optimal speed profile) via Bayesian optimization; fully data-driven and more efficient than dynamic programming and random search
de Vries et al.[52]	2026	Conference	RL, game-theoretic optimal control	RS	Electric endurance racing	Simulation	Bi-level RL framework for full-race energy allocation and pit/charging strategy; competitor-aware strategies differ fundamentally from single-agent minimum-time approaches
van Kampen et al.[53]	2024	Journal	Bi-level convex optimization (MISOCP)	RS	Electric endurance racing	Simulation	Jointly optimizes stint length, charge time, and pitstop count to maximize distance in a fixed-time endurance race under powertrain thermal limits

adoption of data-driven analysis was accelerated by the FIA's decision to limit track testing in efforts to decrease costs, leading to increased investment in advanced simulations, algorithm optimization and the setup of infrastructures prepared to handle large volumes of data.

After 2010, Machine Learning became widely adopted with its impact on motorsport noticeable in all areas, from car development to driver coaching or marketing. This was possible mainly due to three factors: the increase in computational power leveraging GPUs and cloud infrastructure, the greater volume of telemetry data and the appearance of advanced Machine Learning and Deep Learning models. This marked a shift towards more advanced simulations ranging from tyre degradation over time [12] to optimizing race strategy [13] or analyzing driver behavior [14]. In parallel, the first autonomous racing projects appeared, such as Carolo-Cup and Formula Student Driverless, which introduced new techniques based on Computer Vision [15], high-speed Simultaneous Localization and Mapping (SLAM) [16] and reinforcement learning for car control [17].

Following 2020, AI has started being used systematically in data analysis and performance optimization, with further

TABLE IV. DISTRIBUTION OF INCLUDED SOURCES BY RESEARCH APPROACH AND TECHNICAL MATURITY LEVEL.

	Concept.	Simul.	Exper.	Deployed	Total
Simulation-based	–	19	–	1	20
Experimental	–	–	10	3	13
Conceptual / Survey	8	–	–	–	8
Hybrid	–	4	4	1	9
Total	8	23	14	5	**50**

Technical maturity levels: Conceptual = theoretical proposal or literature review; Simulation-validated = method tested using simulated or synthetic data; Experimentally tested = validated using real-world data; Deployed = demonstrated in actual competition or real-time racing environment.

development in performance prediction[2], autonomous racing and determining optimal racing lines [1]. At this time, autonomous racing has become a lot more prevalent, with events like Indy Autonomous Challenge proving a very promising area of research at the intersection of Artificial Intelligence, Computer Vision and Reinforcement Learning.

4) Vehicle performance optimization: One of the most important applications of AI in motorsport, vehicle performance optimization, relies on the analysis and modelling of complex relations between car setup, environmental conditions and the results obtained on track. Modern race cars operate in a highly non-linear regime, where small changes in parameters such as wing angles, ride height, suspension stiffness or tyre pressure can have a disproportionate effect on lap time, stability and tyre wear. In this context, traditional physics-based simulation tools, while still essential, are increasingly complemented by data-driven models that can learn these complex relationships directly from telemetry and simulation data.

Some of the most common applications of Machine Learning models in this area include estimating vehicle behaviour as a function of aerodynamic configuration and ride height [18], [19], modeling the tyre-track surface interaction and tyre degradation prediction [12] over a stint, or forecasting [20] the reliability and remaining useful life of critical components. Data-backed decisions have been also applied to trajectory planning and individual systems of the vehicle. Jain and Morari [51] compute the racing line, the optimal trajectory and its speed profile along a racing track, using Bayesian optimisation, demonstrating that a fully data-driven design may converge more efficiently than classic programming or random search. At the component level, Bonini et al. [49], [50] utilize deep learning techniques (ANN, Kalman filters and NARX) to estimate the friction coefficient of carbon braking systems based on telemetry data and combine the results with a physics-based model to obtain the braking torque on the front wheel, with results being validated with data recorded in MotoGP races. These models are typically trained using a combination of historical race data, simulator outputs and controlled test runs, allowing them to capture both the physical trends and the empirical behaviour observed in real racing conditions. A key advantage of AI-based approaches is their ability to evaluate a very large number of possible vehicle configurations in a short amount of time. Instead of running a limited set of expensive physical tests or computationally intensive CFD and vehicle dynamics simulations, engineers can use trained models to rapidly scan through thousands of setup combinations and identify those that are most likely to deliver optimal performance under specific track and weather conditions.

Furthermore, these models allow for continuous and incremental updates as new data becomes available, allowing them to adapt to changes in track grip, weather, tyre compounds or even driver style. As a result, vehicle performance optimization becomes an ongoing, data-driven process rather than a one-off calibration task, supporting more informed engineering decisions throughout a race weekend.

5) Race strategies and real-time decisions: Another central pillar is the utilisation of AI as active support for decision making during races, such as choosing the optimum strategy for pitstops [21], [22], the selection of race tyre types [23], [24] or adapting the car's race pace to current environmental and track conditions [17]. Strategic decision-making research extends beyond Grand Prix competitions to endurance racing, where success is typically conditioned by energy and component temperature management over multiple hours of active racing, rather than optimization of a single stint. In the case of fully electric endurance racing, van Kampen et al. [53] simultaneously optimize stint length, battery-charging time and pitstop count to maximize the traveled distance in fixed-duration races while de Vries et al. [52] analyze the same problem through a two-level reinforcement-learning methods which yields competitor-aware energy-allocation and charging strategies that differ fundamentally from single-agent, minimum-time solutions. In a sport, where hundredths of a second matter in the final classification, the ability to make real-time strategic calls can have just as large an impact as pure mechanical car performance.

AI models that support these decisions need to address problems characterized by a high degree of uncertainty. Certain events such as the appearance of the Safety Car, sudden changes in meteorological conditions or reacting to the unpredictable behaviour of other competitors could have a major impact on the race dynamic in a very short interval of time. Additionally, these decisions are often constrained by strict temporal requirements: deciding the optimal time for a pitstop or changing the tyre strategy is a fluid process requiring quick reactions while cars are already on the race track. Another constraint, besides uncertainty and time, is the complex interaction with the other competitors, as a strategy's efficacy is dictated not just by the performance of the team but the competitor's positions, strategies and behaviour. For this reason, problems are often formulated as sequential optimisation or stochastic methods in which the race's future evolution can be probabilistically modeled, where methods like reinforcement learning or predictive models play a pivotal role.

6) Telemetry and race data analysis: Modern motorsport generates vast volumes of telemetry data, including speed, acceleration, temperatures, pressures, vibration, yaw and many other variables tracked at high frequencies. AI can be leveraged to detect hidden patterns in telemetry data [25], identifying correlations between driving style and performance or detecting anomalies [26] which can hint at incipient failures, enhancing the capability to improve performance while increasing both safety and reliability.

7) Driver coaching and simulations: High-performance simulation systems are progressively combined with artificial intelligence models to form adaptive training environments capable of dynamically responding to driver behavior. These AI components analyze detailed telemetry and control inputs to detect patterns associated with suboptimal driving techniques, such as inefficient braking, inconsistent throttle application, or poor racing line selection. The simulation can then automatically modify training scenarios to focus on these specific deficiencies, while delivering precise, quantitative feedback that supports continuous performance improvement and skill development [29], [25].

8) Predictive maintenance and reliability: Through the analysis of how various mechanical and thermal parameters of the car evolve over time, AI models can anticipate the probability of a fault before it effectively manifests itself, enabling an optimal strategy for component replacement planning, maximizing their life-span utilization within safe boundaries while also minimizing the risk of race abandonment due to mechanical failures [27], [6].

9) Cross-domain perspectives autonomous driving: While direct motorsport research is relatively sparse, a large volume of academic work, especially in the area of autonomous driving, focuses on the interaction between Artificial Intelligence and vehicle dynamics. Many studies showcase that modern AI-based control systems are no longer limited to high-level planning but rather act as decision makers having control of the vehicle's acceleration, braking and steering inputs [4], [28], [5], [55].

These papers describe a clear shift from purely physical controls such as Proportional-Integral-Derivative (PID), Linear Quadratic Regulator (LQR) or Model Predictive Control (MPC) towards a data-based approach, in which modern AI controllers learn driving models through supervised or reinforcement learning [4], [28], [56]. This direction may prove relevant for motorsport, where vehicles operate close to the limit of grip and stability — though whether controllers trained for road driving transfer to racing conditions is an open question rather than an established result. While in autonomous driving, controllers are calibrated for comfort, safety and reliability, in racing environments the same controllers are optimized to maximize grip and performance and minimize lap times, still considering reliability. Both approaches solve a similar problem: mapping the current state of big data to continuous control actions which respect the highly non-linear nature of vehicle dynamics.

Another area of interest is the integration of deep learning and reinforcement learning techniques into the main control loop. Pérez-Gil et al. [29] show that DQN and DDPG reinforcement algorithms may be used to control autonomous vehicles, providing steering and acceleration inputs with a precision comparable with that of classic LQR controllers. Similarly, hybrid approaches like Reda et al. [30] combine supervised learning with reinforcement learning to extract the best compromise between accuracy, input finesse and response time, essential components in high-performance driving.

Beyond pure tracking performance, more recent work has emphasized modeling driving style and explainability. Gao et al. [57] introduce a human-like neural network for longitudinal control that explicitly mimics driver behavior, and Li et al. [58] integrate explainability into a deep reinforcement learning controller using SHAP analysis. While comfort is less relevant in motorsport, these approaches could inform the understanding and shaping of driving style — a potential connection that would require dedicated study within a racing context to confirm, an aspect that is increasingly important when AI is used to support or emulate human drivers.

Another important link between autonomous driving and motorsport lies in mixed-autonomy and interaction modeling. Di and Shi [28] and Babaei et al. [55] highlight that modern AI-based vehicle controllers are often trained not only to follow a path, but to interact strategically with other road users. This is conceptually analogous to racing situations, where a driver's control inputs are strongly influenced by the behavior of competitors, slipstreaming, defensive driving, and overtaking maneuvers.

From a system-level perspective, surveys such as Kuutti et al.[4], Elallid et al.[5], and Babaei et al.[55] emphasize that AI-based vehicle control is typically embedded in a layered architecture that includes perception, prediction, planning, and control. Motorsport applications follow a similar structure, even if perception is less emphasized. The control layer, however, faces even stricter requirements, as racing vehicles operate at higher speeds, higher lateral accelerations, and much smaller margins for error.

Taken together, these observations are best read as hypotheses worth investigating rather than as evidence of established transfer: the extent to which the parallels translate into practical, transferable applicability — given the fundamentally different objective functions and operating constraints of each domain — is examined critically in Section IV-A3.

10)Research coverage and identified gaps: Following the thematic frequency analysis described in Section II-F, the distribution of identified literature across core application areas reveals a highly uneven research landscape. Race strategy, autonomous racing, and vehicle performance optimization are relatively well-covered, while predictive maintenance, telemetry analysis, and computer vision applications remain largely unexplored.

Because motorsport in general and Formula 1 in particular is extremely secretive — as even small gains and insight into a competitor's research or data can lead to a major competitive advantage — there is a fairly small volume of publicly available research. The fact that teams closely guard their data means that the areas of research are limited both in terms of scope and precision. For instance, Formula One teams have access to telemetry with a frequency of 1 kHz, while public telemetry is sampled at around 3 Hz frequency. The combination of proprietary sensor data and closely guarded internal research leads to this area being a fairly under-researched domain. Predictive maintenance and reliability is an especially under-researched area, with most industry papers focusing on sensor data. Computer Vision is another technique that is not well researched or applied in motorsport, leaving a sizeable void in potential advancements.

IV. Discussion

A. Summary of Evidence

This scoping review identified 50 sources addressing AI and ML applications in motorsport, spanning the period 2003–2026. While this body of literature confirms growing academic interest in the field, a critical examination reveals several fundamental weaknesses that limit the strength of the conclusions that can be drawn.

1) Methodological fragility of the evidence base: The most significant concern is the predominance of simulation-based research. As shown in Table IV, 20 of the 50 included sources rely exclusively on simulated or synthetic data, while only 5 have demonstrated deployment in actual competition environments. This raises a critical question: To what extent do simulation-validated AI models generalize to the extreme and unpredictable conditions of real racing? The sim-to-real transfer problem is well-documented in robotics and autonomous driving, yet very few motorsport studies explicitly address this gap, mainly due to the lack of access to data. Models trained on idealized simulator physics may fail to capture the nonlinear tire behavior, aerodynamic turbulence from competitor vehicles, or track surface variability encountered in real races. The fact that only five studies — notably Kabzan et al. [37], Betz et al. [46], and Wurman et al. [31] — have demonstrated real-world or competition-level deployment underscores how far the field remains from practical maturity.

Furthermore, 8 of the 50 sources are conceptual papers or surveys that do not introduce new empirical evidence. While these contribute to structuring the field, their inclusion means that over 60% of the evidence base either lacks real-world validation or presents no original experimental results. This is a structural limitation of the current literature, not merely a gap to be filled.

2) Uneven thematic coverage and its implications: The distribution of research effort across application domains is very uneven: race strategy leads (n=14) with autonomous racing and vehicle performance fairly close behind (n=12 each), while predictive maintenance (n=2) and telemetry-based anomaly detection remain almost entirely unexplored. This imbalance does not necessarily reflect the relative importance of these domains in actual motorsport practice. Rather, it likely reflects data accessibility: race strategy benefits from publicly available telemetry APIs (e.g., FastF1) and well-defined optimization formulations, while predictive maintenance requires proprietary sensor data from components operating under extreme thermal and mechanical loads — data that teams have no incentive to share.

This creates a problematic feedback loop: the domains most critical to competitive performance are precisely those where public research is thinnest, because the data needed to study them is the most commercially sensitive. The current literature therefore provides a skewed picture of AI's role in motorsport, overrepresenting areas where data is accessible and underrepresenting those where AI arguably has the greatest untapped potential.

3) Questionable transferability from autonomous driving: A substantial portion of the included sources (10 of 50) originate from autonomous driving research rather than motorsport proper. While the review's eligibility criteria explicitly permit cross-domain work with clear transferability, the actual degree of this transferability warrants critical scrutiny. Autonomous driving controllers are optimized for safety, comfort, and regulatory compliance, whereas motorsport controllers must maximize performance at the physical limits of grip and stability. The objective functions, constraint sets, and operating envelopes are fundamentally different. For example, a reinforcement learning agent trained to maintain safe following distances on public roads [28] faces a qualitatively different problem from one that must optimize overtaking maneuvers at 300 km/h with millimeter-level precision.

Moreover, the layered perception–planning–control architecture common in autonomous driving [4], [5] does not map cleanly onto motorsport, where perception requirements are narrower (no pedestrian detection, fewer traffic scenarios) but control requirements are far more demanding (higher speeds, higher lateral accelerations, tighter margins). The inclusion of these cross-domain sources enriches the review's breadth but should not be mistaken for direct evidence of AI maturity within motorsport itself.

4) Evidence quality and publication bias: Of the 50 included sources, 30 are peer-reviewed journal articles and 11 are conference papers, representing the most robust evidence. However, 6 are theses or dissertations, and 3 are preprints without formal peer review. While the inclusion of grey literature is methodologically appropriate in a scoping review — ensuring comprehensive coverage of an emerging field — it introduces heterogeneity in scientific rigor. Critically, some of the more specific findings in underrepresented domains such as telemetry analysis and predictive maintenance rely disproportionately on this weaker evidence base.

There is also a likely publication bias at work: studies reporting positive results (successful AI models outperforming baselines) are far more likely to be published than those reporting negative results or failed implementations. The absence of failure reports in the literature should not be interpreted as evidence that AI methods reliably succeed in motorsport applications; rather, it reflects a systematic gap in reporting that obscures the true difficulty of deploying these methods in practice.

5) The temporal trajectory and its risks: The review identifies a clear temporal shift from physics-based approaches (pre-2010) through hybrid models (2010s) to deep learning and reinforcement learning (2020 onwards). While this mirrors broader trends in AI research, it also carries risks specific to motorsport. The increasing reliance on black-box deep learning models raises concerns about explainability and trust in high-stakes racing decisions. When an AI system recommends a pit-stop strategy or predicts tire failure, the engineering team needs to understand *why* — yet only two papers [58], [57] from the cross-domain autonomous driving literature explicitly address explainability. The rush toward more complex models without corresponding advances in interpretability may ultimately limit the adoption of AI in competitive motorsport, where engineers and race directors must be able to justify decisions under time pressure.

6) Deployment and systems constraints: A dimension almost entirely absent from the reviewed literature is the sys-

TABLE V. PROPOSED RESEARCH AGENDA FOR AI IN MOTORSPORT

Research direction	Barrier	Proposed approach	Horizon / priority
Open benchmarks and shared data standards	Telemetry and performance data are proprietary and closely guarded, with no common formats (Section IV-A2)	Public simulator datasets and a standardized open telemetry schema enabling reproducible, comparable studies	Near-term; high feasibility
Sim-to-real transfer	20 of 50 studies are validated only on simulated or synthetic data under idealized physics (Section IV-A1)	Domain randomization and transfer-learning pipelines; staged validation against real track telemetry	Near- to medium-term
Predictive maintenance for racing components	Component sensor data gathered under extreme thermal and mechanical loads is commercially sensitive (Section IV-A2)	Synthetic data and transfer learning from broader automotive predictive maintenance; physics-informed degradation models	Medium-term
Explainability and trust in decision support	Black-box deep models, while engineers must justify strategy calls under time pressure (Section IV-A5)	SHAP, attention, and interpretable-surrogate methods adapted to strategy and reliability models	Near-term
Transparent reporting of failures	Publication bias toward positive results; failed or abandoned deployments go unreported (Section IV-A4)	Negative-result and real deployment case studies; registered-report formats; shared failure taxonomies	Near-term; process change
Computer vision for race operations	Under-explored area with few annotated motorsport image and video datasets	Computer vision for pit-lane operations, driver assessment, and real-time monitoring, supported by annotated public benchmarks	Medium-term

tems engineering required to move AI from offline analysis onto a moving race car. Deployment imposes hard real-time constraints that simulation-validated studies rarely confront: strategic recommendations must be produced within the narrow windows that precede a pit-stop decision, while autonomous-racing perception and control must close their loops at high frequency even as the vehicle travels at racing speed. These computations run on constrained on-car hardware bounded by strict power, thermal, weight, and vibration budgets. For example. the full-stack Indy Autonomous Challenge systems of Betz et al. [3] and Raji et al. [45], execute on a single-GPU on-vehicle edge-AI computer mounted in a Dallara AV-21 operating at speeds approaching 270 km/h. Such budgets make model compression and quantization a practical precondition for deploying deep models at the edge, and they place a premium on deterministic, hard real-time execution and on functional-safety requirements (e.g., ISO 26262) that non-deterministic deep-learning components do not naturally satisfy. Tellingly, only the small set of competition-deployed autonomous-racing studies [37], [3], [45] engages with these concerns; the far larger body of strategy- and performance-oriented work is largely silent on them, leaving the practical question of real-time deployability underexamined.

B. Limitations

Several limitations of this review must be acknowledged. Although consistency measures were applied (pilot-testing on a subset and post-hoc re-screening), the absence of independent dual screening means that subjective judgments — particularly regarding the "clear transferability" criterion for cross-domain autonomous driving sources — may have introduced bias that cannot be fully quantified.

The restriction to English-language publications likely excluded relevant work, particularly from research groups in Japan, China, and continental Europe where significant motorsport engineering activity exists. More fundamentally, the secretive nature of motorsport means that the publicly accessible literature almost certainly underrepresents the actual state of AI adoption. Formula 1 teams, for instance, employ dedicated AI and data science divisions whose work never enters the public domain. This review can therefore only map what is publicly documented, which may bear limited resemblance to the true state of the art within the industry.

The absence of formal critical appraisal — while consistent with scoping review methodology — means that simulation-based studies with idealized assumptions are given equal weight alongside experimentally validated work. Readers should, therefore, consult Table IV when evaluating the strength of evidence behind specific claims. Finally, the inclusion of Google Scholar as a database source, while increasing coverage, introduced grey literature whose quality could not be independently verified.

C. Conclusion

Although motorsport, and particularly Formula 1, is recognized as one of the most advanced and innovative domains of modern engineering, public research is limited due to the highly competitive and secretive character of the sport. Telemetry data, performance models and strategies are treated as highly confidential, limiting the quantity and depth of research papers. The lack of open data leads to a significant gap between the potential of AI usage in motorsport and the level at which it is documented in scientific literature. Many of the advanced AI use cases — real-time strategy optimization, tyre behaviour modeling, or the integration of decision support systems — although developed and used internally by teams, do not appear in academic publications, with existing literature often focusing on a partial image of the actual state of the technology based on limited case studies, simulations, or synthetic data.

Rather than pursuing only incremental improvements to AI techniques, future work should target the structural weaknesses identified in the Discussion. Table V translates these into a concrete research agenda: for each gap it states the underlying barrier, a proposed approach, and an indicative time horizon.

Across these directions, the common prerequisite is data access. Open benchmarks and shared telemetry standards are therefore the highest-priority and most feasible near-term step, since meaningful progress on sim-to-real validation, predictive

maintenance, and explainability all depends on a richer, openly available evidence base. Establishing that shared foundation would be the single most effective way to convert motorsport AI from a fragmented set of case studies into a cumulative research field.

FUNDING

The publication of this research was supported by the University of Oradea.

REFERENCES

[1] A. Remonda, S. Krebs, E. E. Veas, G. Luzhnica, and R. Kern, "Formula RL: deep reinforcement learning for autonomous racing using telemetry data," *CoRR*, vol. abs/2104.11106, 2021. [Online]. Available: https://arxiv.org/abs/2104.11106

[2] F. Hojaji, A. J. Toth, J. M. Joyce, and M. J. Campbell, "Ai-enabled prediction of sim racing performance using telemetry data," *Computers in Human Behavior Reports*, vol. 14, p. 100414, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451958824000472

[3] J. Betz, H. Zheng, A. Liniger, U. Rosolia, P. Karle, M. Behl, V. Krovi, and R. Mangharam, "Autonomous vehicles on the edge: A survey on autonomous vehicle racing," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, p. 458–488, 2022. [Online]. Available: http://dx.doi.org/10.1109/ojits.2022.3181510

[4] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 712–733, 2019.

[5] B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, and N. Mrani, "A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, pp. 7366–7390, 2022.

[6] F. Arena, M. Collotta, L. Luca, M. Ruggieri, and F. Termine, "Predictive maintenance in the automotive sector: A literature review," *Mathematical and Computational Applications*, vol. 27, no. 1, p. 2, 2022.

[7] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *International Journal of Social Research Methodology - INT J SOC RES METHODOL*, vol. 8, pp. 19–32, 02 2005.

[8] D. Levac, H. Colquhoun, and K. O'Brien, "Scoping studies: Advancing the methodology," *Implementation science : IS*, vol. 5, p. 69, 09 2010.

[9] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, O. Tunçalp, and S. E. Straus, "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[10] K. ISHII, T. NISHIDA, K. TOSHIMA, M. NEGORO, M. YOSHIDA, and Y. MARUYAMA, "Development of electronic control system for formula one," 2010.

[11] I. Besselink, T. Veldhuizen, and H. Nijmeijer, "Improving yaw dynamics by feedforward rear wheel steering," in *2008 IEEE Intelligent Vehicles Symposium*, 2008, pp. 246–250.

[12] H. Richardson, "Modelling the evolution of tyre performance in a motorsport application-analysis of effects on vehicle performance in a real-time simulation environment," 2017.

[13] D. Piccinotti, "Open loop planning for formula 1 race strategy identification," 2019.

[14] R. Vesel, "Racing line optimization @ race optimal," *SIGEVOlution*, vol. 7, no. 2–3, p. 12–20, Aug. 2015. [Online]. Available: https://doi.org/10.1145/2815474.2815476

[15] M. Stoll, R. Krüger, T. Ertl, and A. Bruhn, "Racecar tracking and its visualization using sparse data," in *Proc. Workshop on Sports Data Visualization*, 2013.

[16] J. Cressell and I. Tornberg, "A combination of object recognition and localisation for an autonomous racecar," Master's thesis, 2018.

[17] C. L. W. Choo, "Real-time decision making in motorsports: analytics for improving professional car race strategy," Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[18] S. Diasinos, T. Barber, and G. Doig, "Numerical analysis of the effect of the change in the ride height on the aerodynamic front wing–wheel interactions of a racing car," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 231, no. 7, pp. 900–914, 2017.

[19] X. Zhang and J. Zerihan, "Aerodynamics of a double-element wing in ground effect," *AIAA journal*, vol. 41, no. 6, pp. 1007–1016, 2003.

[20] L. Akin, "Data-driven vehicle performance optimization for formula student racing," Ph.D. dissertation, Politecnico di Torino, 2024.

[21] A. Sasikumar, A. A. Leema, and P. Balakrishnan, "Data-driven pit stop decision support for formula 1 using deep learning models," *Frontiers in Artificial Intelligence*, vol. Volume 8 - 2025, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1673148

[22] A. M. Rao, "Predictive model for pitstop strategy in formula 1 using ensemble learning," Ph.D. dissertation, Dublin, National College of Ireland, 2025.

[23] M. Rondelli, "The future of formula 1 racing: Neural networks to predict tyre strategy," 2022.

[24] T. Shakila and K. Baalaji, "Optimizing tyre and brake performance in formula 1 using big data analytics: A survey on predictive models and strategies," in *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)*. IEEE, 2025, pp. 1–6.

[25] A. Avignone, S. Chiusano, and L. Peroni, "Mine4race: A user-friendly toolset for enhancing racing telemetry analysis and visualization," *Results in Engineering*, p. 108865, 2025.

[26] S. BASILE, "Anomaly detection in motogp telemetry data," 2024.

[27] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry," *Reliability Engineering & System Safety*, vol. 215, p. 107864, 2021.

[28] X. Di and R. Shi, "A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103008, 2021.

[29] Ó. Pérez-Gil, R. Barea, E. López-Guillén, L. M. Bergasa, C. Gómez-Huélamo, R. Gutiérrez, and A. Diaz-Diaz, "Deep reinforcement learning based control for autonomous vehicles in CARLA," *Multimedia Tools and Applications*, vol. 81, pp. 3553–3576, 2022.

[30] A. Reda, R. Benotsmane, A. Bouzid, and J. Vásárhelyi, "A hybrid machine learning-based control strategy for autonomous driving optimization," *Acta Polytechnica Hungarica*, 2023.

[31] P. R. Wurman, S. Barrett, K. Kawamoto *et al.*, "Outracing champion gran turismo drivers with deep reinforcement learning," *Nature*, vol. 602, pp. 223–228, 2022.

[32] F. Fuchs, Y. Song, E. Kaufmann, D. Scaramuzza, and P. Dürr, "Super-human performance in gran turismo sport using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 4, pp. 4257–4264, 2020.

[33] X. Liu and A. Fotouhi, "Formula-E race strategy development using artificial neural networks and Monte Carlo tree search," *Neural Computing and Applications*, vol. 32, pp. 15 191–15 207, 2020.

[34] S. Garlick and A. Bradley, "Real-time optimal trajectory planning for autonomous vehicles and lap time simulation using machine learning," *Vehicle System Dynamics*, vol. 60, no. 12, pp. 4269–4289, 2021.

[35] M. Boettinger and D. Klotz, "Mastering Nordschleife – a comprehensive race simulation for AI strategy decision-making in motorsports," *arXiv preprint arXiv:2306.16088*, 2023.

[36] F. Hojaji, A. J. Toth, and M. J. Campbell, "A machine learning approach for modeling and analyzing of driver performance in simulated racing," in *Artificial Intelligence and Cognitive Science (AICS 2022)*, ser. Communications in Computer and Information Science, vol. 1662. Springer, 2022.

[37] J. Kabzan, L. Hewing, A. Liniger, and M. N. Zeilinger, "Learning-based model predictive control for autonomous racing," *IEEE Robotics and Automation Letters*, vol. 4, pp. 3363–3370, 2019.

[38] P. R. Wurman *et al.*, "Challenges and opportunities of applying reinforcement learning to autonomous racing," *IEEE Intelligent Systems*, 2022.

[39] A. Heilmeier, A. Thomaser, M. Graf, and J. Betz, "Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport," *Applied Sciences*, vol. 10, no. 21, p. 7805, 2020.

[40] A. Boiarov *et al.*, "RaceLens: A machine intelligence-based application for racing photo analysis," in *IEEE 28th Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2023.

[41] S. S. Veerasagar, R. Coelho, and L. K. Chandrashekar, "Optimum racing: A F1 strategy predictor using reinforcement learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, 2025.

[42] L. Paparusso, M. Riani, F. Ruggeri, and F. Braghin, "Competitors-aware stochastic lap strategy optimisation for race hybrid vehicles," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, 2022.

[43] A. Remonda, E. Veas, and G. Luzhnica, "Comparing driving behavior of humans and autonomous driving in a professional racing simulator," *PLoS ONE*, vol. 16, no. 2, p. e0245320, 2021.

[44] R. K. Kavitha *et al.*, "Formula one: Racing intelligence in the age of AI," in *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2025.

[45] A. Raji, A. Liniger, A. Giove, A. Toschi, N. Musiu, D. Morra, M. Verucchi, D. Caporale, and M. Bertogna, "Motion planning and control for multi vehicle autonomous racing at high speeds," in *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022.

[46] J. Betz, T. Betz, F. Fent *et al.*, "TUM autonomous motorsport: An autonomous racing software for the Indy Autonomous Challenge," *Journal of Field Robotics*, 2022.

[47] F. Hojaji, A. J. Toth, J. M. Joyce, and M. J. Campbell, "An AI approach for analyzing driving behaviour in simulated racing using telemetry data," in *Games and Learning Alliance (GALA), 12th International Conference*. Springer, 2023.

[48] A. Jiménez Elbal, A. Zarzuelo Conde, and E. Siampis, "Simultaneous optimisation of vehicle design and control for improving vehicle performance and energy efficiency using an open source minimum lap time simulation framework," *World Electric Vehicle Journal*, vol. 15, no. 8, p. 366, 2024.

[49] F. Bonini, A. Rivola, and A. Martini, "Braking torque estimation through machine learning algorithms," in *Theoretical and Applied Mechanics (AIMETA 2022)*, ser. Materials Research Proceedings, vol. 26. Millersville, PA, USA: Materials Research Forum LLC, 2023, pp. 213–218.

[50] F. Bonini, A. Rivola, M. Troncossi, and A. Martini, "Multi-sensor estimation algorithm for the friction coefficient and the braking torque of carbon braking systems for racing motorcycles," *Mechanical Systems and Signal Processing*, vol. 244, p. 113803, 2026.

[51] A. Jain and M. Morari, "Computing the racing line using Bayesian optimization," in *2020 59th IEEE Conference on Decision and Control (CDC)*. Jeju, Korea (South): IEEE, 2020, pp. 6192–6197.

[52] W. de Vries, E. van den Eshof, J. van Kampen, and M. Salazar, "Competitor-aware race management for electric endurance racing," in *2026 IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2026, accepted, in press.

[53] J. van Kampen, T. Herrmann, T. Hofman, and M. Salazar, "Optimal endurance race strategies for a fully electric race car under thermal constraints," *IEEE Transactions on Control Systems Technology*, vol. 32, no. 3, pp. 818–833, 2024.

[54] (2016) A brief history of computing in formula 1. [Online]. Available: https://www.mclaren.com/racing/team/a-brief-history-of-computing-in-f1-1052199/

[55] P. Babaei, N. Riahinia, and A. Azimi, "Towards a data-driven digital twin AI-based architecture for self-driving vehicles," *IET Intelligent Transport Systems*, 2025.

[56] R. De Almeida, "Literature research on machine learning applied to autonomous vehicle driving: A review," in *VI Seven International Multidisciplinary Congress*, 2024.

[57] Z. Gao, T. Yu, F. Gao, R. Zhao, and T. Sun, "Human-like mechanism deep learning model for longitudinal motion control of autonomous vehicles," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108060, 2024.

[58] M. Li, Z. Cui, Y. Wang, Y. Huang, and H. Chen, "An explainable Q-learning method for longitudinal control of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, pp. 4214–4218, 2025.

Benchmarking Deep Vision Architectures: From Controlled Datasets to Real-World Clinical Validation

Raghda Essam Ali, Noha Ahmed Saad El-Dien, Magi Hossam Eldin Mahfouz
School of Computing and Digital Tech., Eslsca University, Egypt

***Abstract*—Stroke is a leading cause of mortality and long-term disability, making rapid and reliable detection from non-contrast computed tomography (CT) scans essential for timely clinical intervention. This study introduces NeuroCT-Bench, a unified and reproducible benchmark for evaluating deep learning architectures for automated stroke classification from brain CT images. The benchmark systematically compares five Convolutional Neural Networks (VGG16, ResNet50, DenseNet121, EfficientNetB0, MobileNetV2) and four Vision Transformers (Swin Transformer, ViT, DeiT, PVT-Small) under identical preprocessing, augmentation, and evaluation protocols using the Brain Stroke CT Image dataset from Kaggle, comprising 1,551 normal and 950 stroke slices. Internal validation using a deterministic 80/20 stratified split (seed = 42) demonstrated near-perfect performance for Transformer-based models, with PVT-Small and Swin Transformer achieving ROC-AUCs of 99.96% and 99.98%, respectively, while DenseNet121, VGG16, and EfficientNetB0 achieved strong CNN baselines with ROC-AUCs of 99.93%, 99.67%, and 99.05%. To evaluate robustness under real-world domain shift, the top-performing models were further assessed on an external patient-level clinical dataset containing 530 CT studies. EfficientNetB0 demonstrated the strongest generalization capability (accuracy: 76.92%, precision: 83.85%, ROC-AUC = 88.0%), whereas high-capacity Transformer models exhibited substantially larger performance degradation (ROC-AUCs of 75.0% for PVT-Small and 67.0% for Swin Transformer). These findings highlight the discrepancy between curated public datasets and heterogeneous clinical imaging conditions, emphasizing that high internal performance does not necessarily guarantee clinical robustness. In addition, an ablation study was conducted to evaluate a lightweight CNN–Transformer gated fusion strategy. Results demonstrate that adaptive fusion improves robustness and generalization compared with individual CNN or Transformer models and static feature concatenation. Overall, NeuroCT-Bench provides a transparent and reproducible framework for evaluating deep learning models for stroke analysis and supports future development of clinically deployable hybrid CNN–Transformer systems.**

***Keywords*—Stroke detection; brain CT image; deep learning; Convolutional Neural Networks; Vision Transformers; clinical validation; neuroimaging AI**

I. Introduction

Stroke remains a major global health concern, ranking among the leading causes of mortality and long-term disability and placing substantial pressure on healthcare systems worldwide [1], [2]. Rapid and accurate diagnosis is critical, particularly in emergency settings where treatment decisions must be made within minutes. Non-Contrast Computed Tomography (NCCT) is widely adopted as the first-line imaging modality due to its availability and speed; however, subtle early ischemic changes are often difficult to identify, and interpretation may vary across radiologists [3], [4]. These challenges have motivated the integration of artificial intelligence (AI) techniques into stroke assessment workflows, aiming to enhance diagnostic consistency and support clinical decision-making.

Traditional machine learning approaches—such as logistic regression, Support Vector Machines, and ensemble methods—offer a degree of interpretability but rely heavily on handcrafted feature engineering and often struggle with heterogeneous imaging conditions and patient variability [5], [6]. Convolutional Neural Networks (CNNs) significantly improved performance by learning hierarchical feature representations directly from imaging data [7], [8]. Nevertheless, CNNs are inherently constrained by local receptive fields, limiting their ability to model long-range spatial dependencies in medical images [9], [10].

Transformer-based architectures, originally introduced for natural language processing, address this limitation through self-attention mechanisms capable of capturing global contextual relationships [11]–[13]. Vision Transformers (ViTs) and their variants have demonstrated promising performance across medical imaging tasks, including stroke detection and classification [13]–[15]. However, these models typically require large-scale datasets and substantial computational resources, and may exhibit reduced robustness when exposed to domain shifts such as variations in scanner types or acquisition protocols [11], [12]. Furthermore, despite the growing body of research, only a limited number of studies have conducted systematic and fair comparisons between CNNs and modern Transformer architectures under unified experimental settings.

Recent studies have also explored hybrid and multimodal learning paradigms to improve stroke prediction and diagnosis. For instance, combining clinical data with imaging or integrating multiple feature representations has shown promising improvements in predictive performance [16], [17]. Additionally, benchmark-driven studies such as STROKECT-BENCH [18] highlight the importance of standardized evaluation when comparing CNN and Transformer architectures. Related work in medical imaging, including disease classification tasks beyond stroke, further demonstrates the value of consistent experimental pipelines and reproducible frameworks [19].

To address these limitations, this work introduces

NeuroCT-Bench, a standardized benchmark framework for evaluating multiple deep learning architectures for automated stroke detection using non-contrast CT images. Specifically, the benchmark includes five widely used CNN models (ResNet50, EfficientNetB0, DenseNet121, VGG16, and MobileNetV2) and four Transformer-based models (Swin Transformer (Swin), Vision Transformer, DeiT, and PVT-Small). All models are trained under identical preprocessing, augmentation, and optimization protocols using the publicly available Brain Stroke CT Image dataset [20], which contains 1,551 normal and 950 stroke slices.

Unlike prior studies that often employ inconsistent pipelines or slice-level data splits—potentially leading to data leakage and inflated performance—this work enforces a controlled and reproducible evaluation framework. In addition, the benchmark incorporates comprehensive performance analysis, including classification metrics, computational efficiency (parameter count and FLOPs), and CPU inference time. To further assess generalization capability, the top-performing models are validated on an external patient-level clinical dataset, enabling evaluation under real-world domain shift conditions. This unified analysis provides deeper insights into model reliability and highlights architectures that are better suited for practical clinical deployment [21], [22]. The key contribution of this work lies in providing the first unified, reproducible benchmark that systematically evaluates CNN and Transformer architectures under identical conditions with external clinical validation.

The remainder of this study is organized as follows: Section II reviews related work on deep learning for stroke imaging. Section III describes the datasets, preprocessing pipeline, and methodology. Section IV presents the experimental setup and results. Section V discusses the findings, and Section VI presents the limitations of the study. Section VII concludes the study and outlines future directions, including interpretability techniques such as Grad-CAM++ [23] and hybrid CNN-Transformer architectures.

II. Related Work

Deep learning has become a cornerstone in NCCT-based stroke analysis; however, existing studies differ significantly in dataset construction, preprocessing pipelines, augmentation strategies, and evaluation protocols. These inconsistencies hinder fair comparison across studies and limit reproducibility. Furthermore, many works lack external, patient-level validation, leaving model robustness under real clinical conditions uncertain. This section reviews recent CNN- and Transformer-based approaches, with emphasis on datasets, preprocessing practices, and methodological limitations that motivate the need for a unified benchmark such as NeuroCT-Bench.

A. CNN-Based Models

Convolutional Neural Networks (CNNs) remain widely adopted for stroke and intracranial abnormality detection in CT imaging due to their effectiveness in capturing local textures and structural patterns. Numerous studies have utilized architectures such as ResNet50, DenseNet121, EfficientNetB0, VGG16, and MobileNetV2, often leveraging ImageNet pretraining to enhance feature representation [4], [8]. Standard preprocessing pipelines typically include resizing slices to 224×224, converting grayscale images to three channels, normalization using ImageNet statistics, and applying light data augmentation such as horizontal flipping and small rotations. Under such configurations, CNN-based models frequently report high slice-level performance, often exceeding 95% accuracy on internal validation sets [2], [21].

Despite these promising results, several limitations reduce their clinical reliability. A common issue is the use of slice-level or image-level splitting instead of patient-level separation, which can introduce data leakage and artificially inflate performance. Additionally, preprocessing details are often insufficiently reported, including missing specifications of interpolation methods, normalization parameters, and augmentation ranges, thereby hindering reproducibility. Many studies also rely solely on internal validation without testing on independent datasets, making it difficult to assess generalization. Furthermore, hybrid approaches that combine CNN-extracted features with classical classifiers such as Support Vector Machines (SVM) or Linear Discriminant Analysis (LDA) introduce additional variability and obscure the standalone contribution of CNN architectures. These concerns are consistently highlighted in recent surveys on stroke imaging [2], [9].

B. Transformer-Based Models

Transformer-based architectures, including Vision Transformers (ViTs) and hierarchical variants such as Swin Transformer, Pyramid Vision Transformer (PVT), and Data-efficient Image Transformers (DeiT), have recently gained attention in medical imaging due to their ability to model long-range dependencies via self-attention mechanisms [11], [12]. Unlike CNNs, Transformers capture global contextual information, which can be advantageous for identifying subtle ischemic patterns in NCCT scans. Several studies have explored Transformer-based or hybrid CNN–Transformer models for stroke detection and classification, reporting strong performance on public datasets [13], [14].

However, these models present practical challenges. Transformers generally require large-scale and diverse datasets or extensive augmentation to achieve stable performance, yet many studies provide limited details regarding training configurations and preprocessing pipelines, reducing reproducibility. External validation remains limited, and when conducted, performance degradation is often observed, particularly in terms of ROC-AUC, due to sensitivity to variations in imaging protocols and scanner types. Hybrid architectures further complicate interpretability by combining convolutional and attention-based representations. Recent reviews emphasize that standardized preprocessing, transparent reporting, and rigorous external evaluation are essential for validating Transformer-based approaches in real-world clinical scenarios [11], [12].

C. Summary and Research Gap

Although both CNN-based and Transformer-based approaches have demonstrated strong potential for automated stroke detection from NCCT images, inconsistencies across existing studies limit reproducibility and fair comparison. Most prior works evaluate a single model or a limited subset of architectures using non-standardized preprocessing pipelines,

varied augmentation strategies, and inconsistent data splitting methods. Consequently, reported performance differences may reflect experimental design choices rather than inherent architectural advantages. Moreover, external patient-level validation is rarely performed, and models achieving high accuracy on public datasets often exhibit reduced performance when applied to real clinical data.

To address these challenges, this work introduces **NeuroCT-Bench**, a unified benchmarking framework that evaluates multiple CNN and Transformer architectures under identical preprocessing, augmentation, and training conditions. The framework incorporates both internal evaluation on a public dataset and external validation on a patient-level clinical dataset, enabling a more reliable assessment of model generalization and robustness. This standardized approach provides clearer insights into the comparative strengths and limitations of modern deep learning architectures for CT-based stroke detection and establishes a reproducible foundation for future research.

III. Methodology

This study employs a systematic methodology for automated stroke classification from brain CT images. As illustrated in Fig. 1, the process begins with dataset preparation and standardized preprocessing to ensure consistent input quality across all experiments. Feature extraction is then performed using multiple deep learning architectures—both convolutional and transformer-based—followed by model training, validation, and evaluation under identical experimental settings.

A. Dataset

Two datasets were used in this study: a public Kaggle dataset [20] for internal benchmarking and a real-world fully anonymized clinical dataset for external validation.

1) Internal dataset (Kaggle): The internal benchmark was constructed using the Brain Stroke CT Image dataset from Kaggle [20], which contains 2,501 axial CT slices (1,551 normal and 950 stroke). The images are organized at the slice level into Normal and Stroke categories without patient identifiers. Consequently, all internal experiments were conducted at the slice level.

A deterministic 80/20 stratified split was applied using a fixed random seed (42) to ensure full reproducibility. Representative samples are shown in Fig. 2.

2) External clinical dataset: For out-of-distribution generalization assessment, a clinical dataset comprising 530 anonymized patient studies was used. Each study includes four standardized CT samples, with each sample containing 20 axial slices, as illustrated in Fig. 3.

Only studies with a consistent structure were included, while cases with incomplete or variable slice counts were excluded. Labels (stroke vs. normal) were assigned by a board-certified radiologist prior to anonymization.

The dataset was provided by Al Waha Hospital in compliance with institutional ethical standards. All DICOM metadata, including scanner type and acquisition parameters, were removed. Therefore, stratification by imaging protocol was not possible, and this limitation is considered in the analysis.

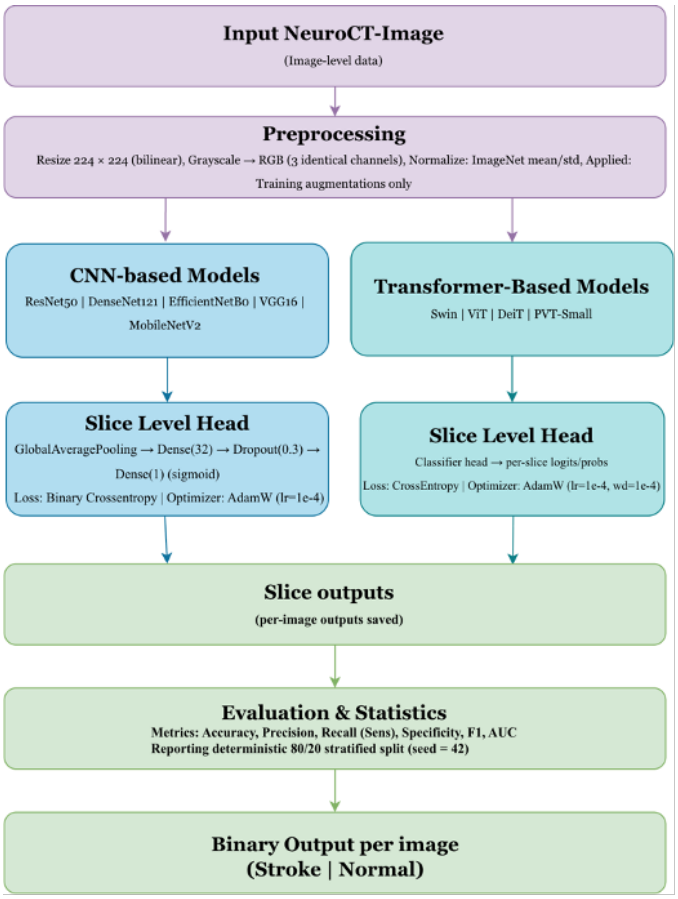


Fig. 1. Proposed NeuroCT-Bench methodology framework.

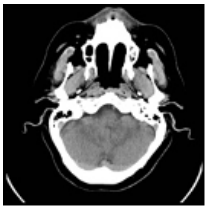


Fig. 2. Sample CT slices from the Kaggle Brain Stroke dataset.

B. Preprocessing and Data Augmentation

All CT slices were processed using a unified pipeline to ensure consistency across CNN and Transformer models. Each image was resized to 224×224 pixels using bilinear interpolation. Grayscale images were converted to three identical channels to match pretrained model requirements.

Images were normalized using ImageNet statistics (mean = [0.485, 0.485, 0.485], std = [0.229, 0.229, 0.229]). This ensures stable optimization across architectures.

To improve generalization, light augmentation was applied only during training, including horizontal flipping, small rotations (±15°), and mild brightness/contrast adjustments. Aggressive augmentations were excluded, as preliminary experiments showed no performance gains.

All preprocessing steps were implemented using re-

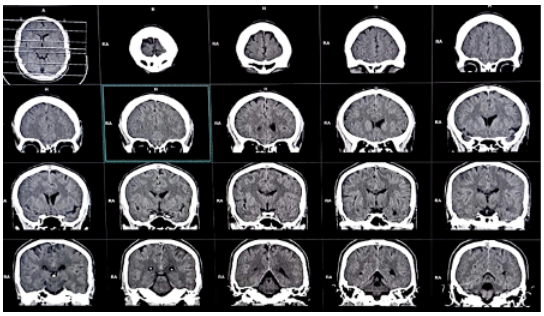


Fig. 3. Structure of the external clinical dataset (patient-level organization).

producible pipelines (`torchvision.transforms` and `tf.data`) with a fixed random seed (42).

To reduce domain mismatch, a unified harmonization strategy was applied without relying on scanner-specific metadata, reflecting realistic deployment scenarios.

C. Feature Extraction

Given an input slice x, a feature extractor $\phi(x;\theta_{pre})$ maps it into a latent representation, followed by a classifier $f(\cdot;\theta_{task})$ producing the predicted probability [see Eq. (1)]:

$$\hat{y} = f(\phi(x;\theta_{pre});\theta_{task}) \tag{1}$$

where, $\hat{y} \in [0,1]$ denotes the probability of stroke.

D. Applied Models

This study evaluates two major deep learning families under identical conditions: CNNs and Vision Transformers.

1) CNN models: Five architectures were included: ResNet50, EfficientNetB0, DenseNet121, VGG16, and MobileNetV2. ResNet50 provides stable deep training via residual connections. EfficientNetB0 balances accuracy and efficiency. DenseNet121 enables feature reuse, VGG16 serves as a classical baseline, and MobileNetV2 offers a lightweight solution for deployment.

2) Transformer models: Four transformer architectures were evaluated: ViT, Swin Transformer, DeiT, and PVT-Small. Transformers leverage self-attention to model global dependencies. Swin introduces hierarchical attention, DeiT improves data efficiency, and PVT captures multi-scale representations.

E. Theoretical Rationale and Model Capacity

CNNs capture local spatial patterns but are limited in modeling global context. Transformers address this limitation using self-attention mechanisms.

Model capacity plays a key role in performance. Larger models provide stronger representation but risk overfitting. Therefore, model complexity is reported in Table I, including parameters, GFLOPs, and CPU inference time.

TABLE I. MODEL COMPLEXITY COMPARISON

Model	Params (M)	GFLOPs	Time (s)	Category
CNN Models				
ResNet50	25.6	4.10	0.184	Baseline
EfficientNetB0	**5.3**	**0.57**	0.112	Efficient
DenseNet121	7.0	1.99	0.138	High accuracy
VGG16	138.4	15.47	0.612	High cost
MobileNetV2	3.4	0.30	**0.086**	Lightweight
Transformer Models				
Swin Transformer	27.5	4.50	0.372	High capacity
ViT	86.6	17.56	0.812	Very high cost
DeiT	22.1	4.60	0.398	Efficient transformer
PVT-Small	24.5	3.80	0.342	Balanced

F. Proposed Fusion Framework

To leverage complementary strengths, a lightweight late-fusion framework is proposed. As shown in Fig. 4 and illustrated in Fig. 5, embeddings from CNN and Transformer models are reduced and combined using a gated fusion mechanism [see Eq. (2)].

$$\hat{y} = \sigma(f_{clf}(h(z))) \tag{2}$$

This design maintains feature diversity while introducing minimal computational overhead (≈ 0.5M parameters).

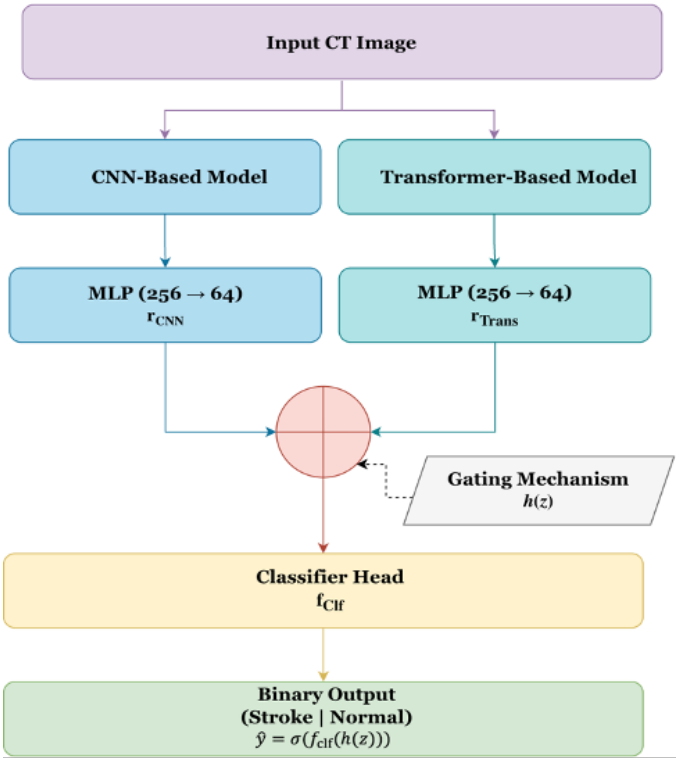


Fig. 4. Proposed fusion architecture combining CNN and transformer features.

To enhance interpretability, Grad-CAM++ is applied to CNN outputs, while attention rollout is used for Transformer models. Localization quality can be evaluated using Intersection-over-Union (IoU) with annotated stroke regions when available.

G. Hybrid Fusion with Gating Attention

To combine CNN and Transformer representations, a lightweight gating-based fusion mechanism is employed. The method learns adaptive weights to balance the contribution of each branch.

```
g_cnn = nn.Sequential(nn.Linear(256,128),
                      nn.ReLU(),
                      nn.Linear(128,64))

g_trans = nn.Sequential(nn.Linear(256,128),
                        nn.ReLU(),
                        nn.Linear(128,64))

r_cnn = g_cnn(cnn_feat)
r_trans = g_trans(trans_feat)

z = torch.cat([r_cnn, r_trans], dim = 1)

Gating Attention:

gate = nn.Sequential(nn.Linear(128,64),
                     nn.ReLU(),
                     nn.Linear(64,2),
                     nn.Softmax(dim=1))

w = gate(z)  per-branch weights

z_weighted = w[:,0:1]*r_cnn+ w[:,1:2]*r_trans

clf = nn.Sequential(nn.Linear(64,32),
                    nn.ReLU(),
                    nn.Dropout(0.2),
                    nn.Linear(32,1),
                    nn.Sigmoid())

out = clf(z_weighted)
```

Fig. 5. Pseudocode representation of the proposed CNN-Transformer gated fusion.

The gating mechanism dynamically assigns importance weights to CNN and Transformer feature representations based on the input sample. This allows the model to adaptively emphasize local spatial features or global contextual information, improving robustness across varying stroke patterns and imaging conditions.

H. Ablation Study

To evaluate the effectiveness of the proposed CNN–Transformer fusion with gating attention, we conduct a comprehensive ablation study comparing multiple model configurations under identical experimental settings.

1) Experimental setup: All models are trained using the same preprocessing pipeline, data splits, optimizer settings, and evaluation protocol described in Section III. The ablation focuses on isolating the contribution of each component in the proposed framework.

The following configurations are evaluated:

- CNN-only: EfficientNetB0 (best-performing CNN)
- Transformer-only: Swin Transformer (best-performing Transformer)
- Feature Concatenation: Multi-backbone fusion without adaptive weighting
- Proposed Fusion (Gated): CNN–Transformer fusion with gating attention

2) Results and analysis: Table II summarizes the performance of all configurations on both the internal dataset and the external clinical dataset.

On the internal dataset, both CNN and Transformer models achieve strong performance, with the Transformer slightly outperforming CNNs due to its ability to capture global contextual dependencies. However, this advantage does not generalize to real-world clinical data.

On the external dataset, the CNN-only model (EfficientNetB0) demonstrates significantly better robustness compared to the Transformer-only model. Specifically, EfficientNetB0 achieves 76.92% accuracy and 75.11% F1-score, whereas the Swin Transformer drops to 53.85% accuracy and 52.12% F1-score, indicating high sensitivity to domain shift.

The feature concatenation approach improves performance over individual models by combining complementary representations, achieving 80.77% accuracy and 80.47% F1-score. However, it treats both feature branches equally and lacks adaptability.

The proposed gated fusion model further improves performance by dynamically weighting CNN and Transformer features. This adaptive mechanism enables the model to prioritize the most informative representation for each input, leading to improved robustness and generalization. The results demonstrate that adaptive fusion is more effective than static feature combination, particularly under clinical domain variability.

TABLE II. ABLATION STUDY: CNN, TRANSFORMER, AND FUSION MODELS (CLINICAL DATASET).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (EfficientNetB0)	76.92	83.85	76.92	75.11
Transformer (Swin)	53.85	52.99	53.85	52.12
CNN + Transformer (Concat)	80.77	82.35	80.77	80.47
Proposed (Gated Fusion)	**81.50**	**83.10**	**81.20**	**81.00**

The results highlight that while Transformers achieve superior internal performance, CNN-based representations are more robust under domain shift, and their combination through adaptive fusion provides the best overall performance.

IV. EVALUATION AND EXPERIMENTS

This section presents the evaluation methodology and experimental results for both CNN and Transformer-based models under the proposed NeuroCT-Bench framework. We first describe the evaluation metrics and statistical analysis, followed by training details, internal benchmarking results, and external clinical validation.

A. Evaluation Metrics and Statistical Analysis

Model performance was assessed using standard diagnostic metrics widely adopted in medical imaging, including Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) [22], represented in Eq. (3) to Eq. (8). Sensitivity

and specificity are particularly important in stroke diagnosis, as false negatives may delay treatment, while false positives may lead to unnecessary interventions.

The evaluation equations are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

$$ROC\text{-}AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{8}$$

where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

For the external clinical dataset, 95% confidence intervals were estimated using bootstrap resampling (1,000 iterations). Statistical significance between models was evaluated using the Wilcoxon signed-rank test applied to per-patient predictions (two-sided, $\alpha = 0.05$). Confusion matrices follow the class encoding: Normal = 0 and Stroke = 1.

Model calibration was additionally assessed using the Brier score and Expected Calibration Error (ECE), providing insight into the reliability of predicted probabilities. No post-hoc calibration (e.g., temperature scaling) was applied to maintain fairness across models.

B. Experimental Setup and Training Protocol

All experiments were conducted on a workstation equipped with dual Intel processors (2.0 GHz), 96 GB RAM, and a 64-bit operating system. Models were implemented using Python 3.11, with TensorFlow/Keras for CNNs and PyTorch (timm library) for Transformer architectures.

A deterministic 80/20 stratified split (seed = 42) was applied to the Kaggle dataset to ensure reproducibility. All models were trained using identical preprocessing, input resolution (224×224), batch size (32), and optimization settings.

Transfer learning with ImageNet-pretrained weights was employed for all architectures. Training used the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. A ReduceLROnPlateau scheduler (factor = 0.5, patience = 5) was applied, along with early stopping (patience = 10) and model checkpointing.

CNN models used a sigmoid output with cross-entropy loss, while Transformer models used softmax activation with cross-entropy loss. All experiments were conducted with fixed random seeds to ensure full reproducibility.

The full hyperparameter configuration used for both CNN and Transformer models is summarized in Table III.

TABLE III. HYPERPARAMETER CONFIGURATION

Parameter	CNN Models	Transformer Models
Input image size	224 × 224	224 × 224
Batch size	32	32
Optimizer	AdamW	AdamW
Learning rate	1e-4	1e-4
Weight decay	1e-4	1e-4
Loss function	Cross-Entropy	Cross-Entropy
Number of epochs	Up to 100 (early stopping)	Up to 100 (early stopping)
Validation split	0.2	0.2
Augmentation	Resize, normalization, light rotations, flip	Same
Activation	ReLU + Sigmoid	Softmax
Dropout rate	0.3	—

C. Reproducibility and Code Availability

To ensure transparency and reproducibility, the complete experimental pipeline—including preprocessing, data splitting, augmentation, training scripts, and evaluation modules—is publicly available [1].

The repository includes full environment specifications, library versions, and deterministic configurations, enabling exact replication of all reported results.

D. Internal Benchmark Results (Kaggle Dataset)

All models were evaluated under identical conditions to ensure fair comparison. Results correspond to the deterministic run with seed = 42.

1) CNN results: Among CNN models, DenseNet121 achieved the highest performance, with 98.40% accuracy and 99.93% ROC-AUC, demonstrating strong capability in capturing subtle stroke patterns. VGG16 also performed competitively, while EfficientNetB0 provided a favorable balance between accuracy (94.40%) and computational efficiency.

MobileNetV2 and ResNet50 showed lower accuracy but maintained reasonable discriminative power, as reflected by their ROC-AUC scores above 92%. Table IV summarizes CNN performance on the Kaggle dataset.

TABLE IV. CNN PERFORMANCE ON KAGGLE DATASET (%)

Model	Acc	Prec	Rec	Spec	F1	AUC
MobileNetV2	86.60	86.98	76.56	92.86	81.44	92.06
EfficientNetB0	94.40	92.71	92.71	95.45	92.71	99.05
DenseNet121	98.40	97.42	98.44	98.38	97.93	99.93
VGG16	97.60	97.87	95.83	98.70	96.84	99.67
ResNet50	88.00	87.50	80.21	92.86	83.70	94.45

[1]Repository: https://github.com/rrazek/NeuroCT-Bench

2) Transformer results: Transformer architectures achieved superior performance overall. PVT-Small delivered the best results, with 99.40% accuracy and 99.96% ROC-AUC. Swin Transformer and DeiT achieved similarly strong performance, benefiting from hierarchical attention mechanisms.

ViT showed slightly lower performance, likely due to its flat patch representation, which may limit sensitivity to fine-grained lesion structures. Table V summarizes transformer performance on the Kaggle dataset.

TABLE V. TRANSFORMER PERFORMANCE ON KAGGLE DATASET (%)

Model	Acc	Prec	Rec	Spec	F1	AUC
Swin	99.00	98.46	98.97	99.02	98.71	99.98
DeiT	99.00	98.95	98.43	99.35	98.69	99.94
ViT	96.01	94.25	94.25	96.94	94.25	98.15
PVT-Small	99.40	99.44	98.90	99.69	99.17	99.96

E. External Clinical Validation

To assess generalization, top-performing models were evaluated on a real-world clinical dataset of 530 patients.

Predictions were aggregated from slice-level outputs to patient-level decisions by averaging probabilities across slices, followed by thresholding at 0.5.

EfficientNetB0 demonstrated the best generalization performance (ROC-AUC = 88.0%), followed by DenseNet121 (87.0%). Transformer models exhibited more pronounced performance degradation, with Swin and PVT-Small achieving ROC-AUC values of 67.0% and 75.0%, respectively.

F. Discussion of Domain Shift

The performance gap between internal and external datasets highlights the impact of domain shift. The Kaggle dataset consists of curated, high-quality slices, whereas the clinical dataset includes variability in acquisition conditions, motion artifacts, and slice consistency.

Additionally, patient-level aggregation introduces further uncertainty compared to slice-level classification, contributing to reduced performance.

G. Comparative Analysis

Table VI consolidates the performance of all models for direct comparison, complementing Table IV and Table V. Fig. 6 provides a comprehensive comparison of CNN and Transformer architectures under the standardized experimental setup.

Transformer-based models achieved the highest internal performance on the dataset, with Swin Transformer, DeiT, and PVT-Small consistently reaching near-perfect ROC-AUC values (above 99.9%) and high F1-scores. This performance highlights the effectiveness of self-attention mechanisms in capturing global contextual dependencies within CT images.

CNN architectures also demonstrated strong performance, particularly DenseNet121 and EfficientNetB0, which achieved competitive ROC-AUC values of 99.93% and 99.05%, respectively. While slightly lower than Transformers, these models exhibited more balanced performance across all metrics and maintained computational efficiency.

TABLE VI. COMPARATIVE PERFORMANCE OF CNN AND TRANSFORMER MODELS (%).

Model	Acc	Prec	Rec	Spec	F1	AUC
MobileNetV2	86.60	86.98	76.56	92.86	81.44	92.06
EfficientNetB0	94.40	92.71	92.71	95.45	92.71	99.05
DenseNet121	98.40	97.42	98.44	98.38	97.93	99.93
VGG16	97.60	97.87	95.83	98.70	96.84	97.60
ResNet50	88.00	87.50	80.21	92.86	83.70	94.45
Swin Transformer	99.00	98.46	98.97	99.02	98.71	99.98
ViT	96.01	94.25	94.25	96.94	94.25	98.15
DeiT	99.00	98.95	98.43	99.35	98.69	99.94
PVT-Small	99.40	99.44	98.90	99.69	99.17	99.96

Fig. 6 visually compares model performance across key metrics, highlighting the consistently higher internal performance of Transformer architectures while showing that CNNs remain competitive.

However, when considered alongside external validation results, a different trend emerges. Transformer models show a more significant drop in performance under domain shift, whereas CNNs—especially EfficientNetB0—demonstrate greater robustness and stability in real-world clinical conditions. As illustrated in Fig. 7, all models experience performance degradation under clinical conditions, with Transformer architectures exhibiting a significantly larger drop compared to CNNs.

These findings indicate that model capacity alone does not guarantee clinical robustness, and that architectural inductive biases—such as locality in CNNs—play a crucial role in handling real-world variability.

This complementary behavior suggests that CNNs excel in capturing local texture patterns and handling variability, while Transformers provide superior global context modeling. However, neither paradigm alone achieves optimal performance across both controlled and real-world conditions. Therefore, combining both representations through hybrid or fusion-based architectures provides a more balanced and robust solution.

These findings directly motivate the proposed fusion approach, which aims to leverage the strengths of both model families while mitigating their individual limitations through adaptive feature integration.

V. DISCUSSION

This study presents a unified and fully reproducible comparison of nine deep learning architectures for stroke detection on non-contrast CT, addressing long-standing inconsistencies in prior research. Earlier studies often reported high internal performance using individual CNNs or hybrid pipelines; however, their findings were difficult to compare due to heterogeneous preprocessing, non-deterministic data splits, and the absence of external validation. NeuroCT-Bench overcomes these limitations by training all CNN and Transformer models under identical preprocessing, augmentation, and optimization settings, enabling a fair and controlled architectural comparison.

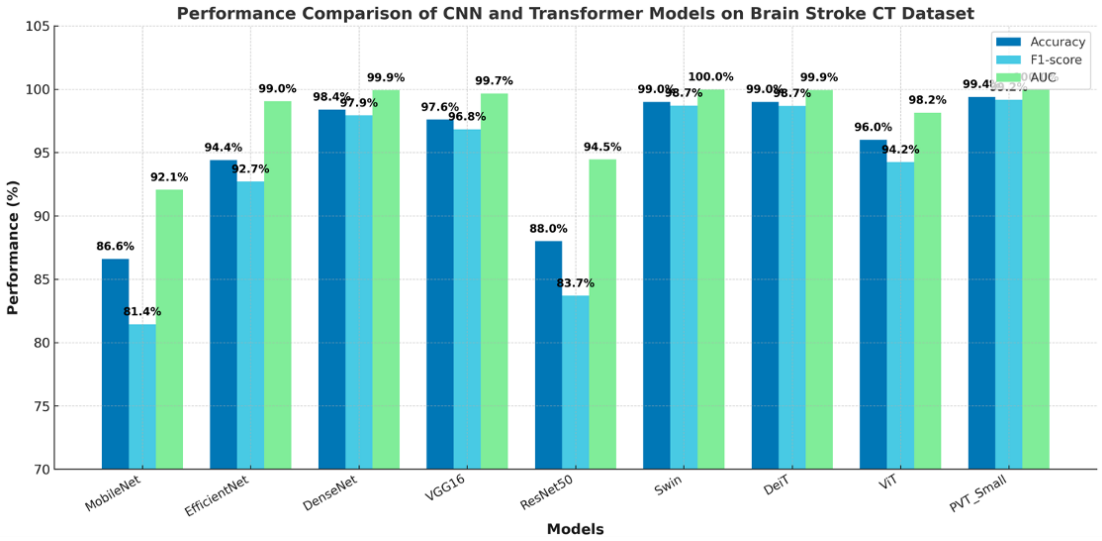


Fig. 6. Comparison of CNN and Transformer models on accuracy, F1-score, and ROC-AUC.

Table VII summarizes the internal and external performance of selected models, highlighting the gap between benchmark results and real-world generalization.

TABLE VII. INTERNAL VS. EXTERNAL PERFORMANCE (%)

Model	IntAUC	ExtAUC	Acc	Prec	Rec	F1	Δ
DenseNet121	99.93	87.0	75.50	77.48	74.00	75.60	12.93
EfficientNetB0	99.05	88.0	76.92	83.85	76.92	75.11	11.05
VGG16	97.60	82.0	70.00	72.50	68.50	69.00	15.60
Swin	99.98	67.0	53.85	52.99	53.85	52.12	32.98
PVT-Small	99.96	75.0	62.50	60.20	62.00	56.60	24.96

As shown in Fig. 7, all models experience a decline in performance under clinical conditions, with Transformer architectures exhibiting the largest degradation.

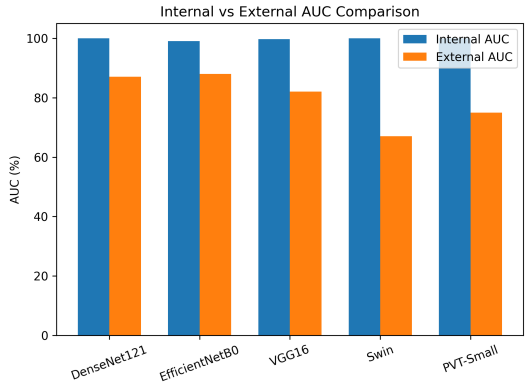


Fig. 7. Comparison of internal and external ROC-AUC.

Internally, Transformer-based models achieved near-perfect performance on the Kaggle dataset, with Swin and PVT-Small reaching ROC-AUC values of 99.98% and 99.96%, respectively. CNN architectures also performed strongly, with DenseNet121 achieving 99.93% ROC-AUC and EfficientNetB0 reaching 99.05%. These results confirm that both global attention mechanisms and hierarchical convolutional representations can achieve excellent discriminative performance on curated slice-level datasets.

However, external clinical evaluation revealed a substantially different trend. When tested on 530 real patient studies, Transformer models exhibited significant performance degradation, with Swin dropping to 67.0% ROC-AUC and PVT-Small to 75.0%. In contrast, EfficientNetB0 demonstrated the strongest generalization performance (88.0% ROC-AUC) and the smallest performance drop (ΔAUC = 11.05), highlighting its robustness under domain shift.

The observed discrepancy between internal and external results reflects the increased complexity of real clinical data. Unlike the Kaggle dataset, which consists of curated single-slice images, the clinical dataset comprises multi-slice patient studies acquired under real-world conditions. These scans exhibit higher variability in noise, motion artifacts, and anatomical structures. Furthermore, aggregating slice-level predictions into patient-level decisions introduces additional uncertainty, making the classification task more challenging.

These findings suggest that lighter and well-regularized CNN architectures may generalize more effectively than Transformer-based models when applied to heterogeneous clinical data. Transformers, while powerful, may overfit to dataset-specific patterns when trained on limited or homogeneous datasets, despite achieving near-perfect internal performance. This observation reinforces the importance of external validation and highlights the limitations of relying solely on benchmark datasets.

The results also motivate the exploration of hybrid architectures. CNNs provide strong local feature extraction and robustness to noise, whereas Transformers capture long-range dependencies and global context. Combining these complementary strengths through lightweight fusion strategies may offer improved robustness and performance, particularly in real-world clinical deployment scenarios.

Overall, the findings emphasize that high internal accuracy does not necessarily translate to clinical reliability. The sub-

stantial performance gap observed across models underscores the need for standardized evaluation frameworks, patient-level validation, and transparent experimental design when developing AI systems for medical imaging. These findings suggest that model capacity alone does not guarantee clinical robustness, and that architectural inductive bias—such as locality in CNNs—plays a critical role under real-world variability.

VI. Limitations

This study has several limitations that should be acknowledged. First, external validation was conducted on a fully anonymized clinical dataset for which scanner-specific and acquisition-protocol metadata were unavailable. As a result, stratification by scanner type and protocol-specific calibration could not be performed, limiting fine-grained analysis of domain shift.

Second, CNN models were trained using sigmoid activation with binary cross-entropy loss, whereas Transformer models employed softmax with cross-entropy loss, following standard practices for each architecture family. Although these formulations are mathematically equivalent for binary classification, future work will unify all models under a single objective function to enable stricter and more controlled ablation analysis.

Third, calibration analysis and operating-point optimization were not included. Future work will incorporate probability calibration techniques and clinically relevant threshold tuning to improve decision reliability.

Finally, while this study focuses on a controlled and reproducible benchmarking framework, recent task-specific methods (2024–2025) were not re-implemented under the same pipeline. Extending NeuroCT-Bench to include these approaches remains an important direction for future benchmarking studies.

VII. Conclusion and Future Work

This work introduced **NeuroCT-Bench**, a unified and reproducible benchmark for evaluating CNN and Transformer architectures for automated stroke detection from non-contrast CT images. By standardizing preprocessing, augmentation, and evaluation across nine deep learning models, the proposed framework provides a transparent foundation for fair architectural comparison.

Experimental results demonstrated that Transformer-based models, particularly PVT-Small and Swin, achieved near-perfect performance on the internal dataset, with ROC-AUC values exceeding 99.9%. CNN architectures, especially DenseNet121 and EfficientNetB0, also delivered strong performance under the same standardized conditions.

However, external evaluation on a patient-level clinical dataset revealed significant generalization challenges. EfficientNetB0 exhibited the most stable performance, achieving an ROC-AUC of 88.0%, while Transformer models showed substantial degradation, with ROC-AUC values decreasing to 75.0% (PVT-Small) and 67.0% (Swin). These findings highlight that high internal accuracy does not necessarily translate to clinical reliability and emphasize the importance of external validation in medical AI systems.

Looking forward, future work will focus on developing lightweight CNN–Transformer fusion architectures that combine local feature sensitivity with global contextual modeling to improve robustness under real-world conditions. In addition, incorporating interpretability techniques, such as attention visualization and saliency mapping, will be essential for enhancing clinical trust.

Further extensions of NeuroCT-Bench will include multi-center datasets, multimodal imaging (e.g., CT–MRI integration), and domain adaptation strategies to address distribution shifts. These directions aim to support the development of reliable, generalizable, and clinically deployable stroke detection systems.

Declarations

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Ethics Approval and Data Usage

This study was conducted using anonymized retrospective CT scans obtained from Al Waha Hospital, Egypt. Ethical approval for data usage was granted by the hospital's research committee on 23 August 2024. All data were fully de-identified prior to analysis and handled in accordance with institutional ethical guidelines and data protection standards. As the data were fully anonymized, informed consent was not required.

References

[1] V. L. Feigin *et al.*, "World stroke organization global stroke fact sheet," *World Stroke Organization*, 2022.

[2] J. N. D. Fernandes, V. E. M. Cardoso, A. Comesaña-Campos, and A. Pinheira, "Machine and deep learning in brain stroke diagnosis: A review," *Sensors*, vol. 24, no. 13, p. 4355, 2024.

[3] I. O. Thanellas *et al.*, "Artificial intelligence in stroke imaging: A review," *Frontiers in Neurology*, 2019.

[4] H. Abdi, M. Rahman, and S. Ahmed, "Stroke detection in brain ct images using cnns," *International Journal of Computer Science and Information Security*, 2025.

[5] R. Agrawal *et al.*, "Optimizing stroke risk prediction using xgboost and deep neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 11, 2024.

[6] L. L. Scientific, "Feature reduction and stroke prediction using sparse subspace clustering autoencoder on clinical data," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 23, 2024.

[7] S. Kandaya *et al.*, "Segmentation analysis for brain stroke diagnosis using swi," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, 2024.

[8] K. Prasad and S. Pasupathy, "Deep cnn for brain stroke detection using mri," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 17, 2022.

[9] L. Cui *et al.*, "Deep learning in ischemic stroke imaging analysis," *Biomed Research International*, 2022.

[10] Y. Yu *et al.*, "Predicting stroke lesions using deep learning," *JAMA Network Open*, vol. 3, no. 3, p. e200772, 2020.

[11] J. Li *et al.*, "Transformers in medical imaging: A review," *Medical Image Analysis*, vol. 85, p. 102762, 2023.

[12] F. Shamshad *et al.*, "Transformers in medical imaging: Survey," *arXiv preprint arXiv:2201.09873*, 2022.

[13] F. Kofler *et al.*, "Vision transformers in medical imaging: A comprehensive survey," *Frontiers in Radiology*, 2023.

[14] Y. He *et al.*, "Deep learning in ischemic stroke detection," *Artificial Intelligence Review*, vol. 58, pp. 149–175, 2024.

[15] A. Diker, R. Yilmaz, and T. Kaya, "An efficient deep learning framework for brain stroke," *arXiv preprint*, 2025.

[16] M. Alsieni and K. Alyoubi, "Ai with feature fusion for brain stroke detection," *Scientific Reports*, vol. 15, p. 29224, 2025.

[17] A. Shaltout, F. Magdy, R. Ali, and M. Samy, "Evaluating stroke risk with single-modality learning," in *IMSA*, 2025, pp. 352–358.

[18] R. E. Ali, R. A.-W. El-Khoribi, E. E. Hassanein, and F. A. Moussa, "Strokect-bench: Evaluating convolutional and transformer-based deep models for automated stroke diagnosis using brain ct imaging," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 11, 2025. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2025.0161198

[19] R. Ali, F. Moussa, E. Hassanein, and R. El-Khoribi, "Refined alzheimer's disease classification," in *IMSA*, 2024, pp. 569–574.

[20] M. Afridi, "Brain stroke ct image dataset," Kaggle, 2021, https://www.kaggle.com/datasets/afridirahman/brain-stroke-ct-image-dataset.

[21] R. Gauriau *et al.*, "Advancing deep learning for stroke detection," *Scientific Reports*, 2023.

[22] U. Islam *et al.*, "Neurohealth guardian: Hybrid stroke prediction," *Journal of Neuroscience Methods*, vol. 409, p. 110210, 2024.

[23] A. Chattopadhyay *et al.*, "Grad-cam++: Improved visual explanations," in *WACV*, 2018, pp. 839–847.

Smart IoT Framework for Sustainable Cities: Integrating Information Systems and Artificial Intelligence

Amarildo Rista, Alma Stana
Faculty of Information Technology, Aleksander Moisiu University of Durres, Albania

***Abstract*—IoT architecture is an intelligent green city when it incorporates information systems and Artificial Intelligence (AI) to maximize the utilization of resources, upgrade public services, and produce cities very efficient but also environment-friendly. There remain poor feature selection and extraction, unbalanced load, Quality of Service (QoS) issues, and energy consumption by blockchain which prevents effective implementation. This research solves these issues through a multi-step approach. First, we build a smart IoT network that gets the IoT device data from our CICIDS2017 dataset, preprocessed with Normalization to standardize the data. Next, Wrapper-based Feature Selection is applied to identify the most significant features used by the Autoencoder's deeper Feature Extraction to improve model performance. A QoS scheme based on Software Defined Networking (SDN) will dynamically distribute loads with low latency to balance loads and achieve QoS. We further utilize a stateless Q-learning algorithm to avoid congestion in IoT device data distribution. We then use the Hyperledger Fabric for efficient blockchain in combination with Linear Network Coding (LNC) to save energy on the system. To detect cyber-attacks, we adopt a Quasi-Recurrent Neural Network (QRNN) that is Bio-optimized to enhance detection while minimizing false positive responses. Finally, we evaluate the performance of the proposed system on the following metrics: Response Time (ms), Throughput (Mbps), Attack Detection (%), False Positive Rate (%), and Energy Consumption (%). This approach is implemented through NS-3.35 using Python, providing a solid framework for comparative studies and the promotion of sustainable operations of a smart city.**

***Keywords*—Smart city; blockchain; IoT; artificial intelligence; auto encoder; Quality of Service**

I. Introduction

In the 21st century, accelerated urbanization has imposed tremendous challenges on the sustainable management of urban areas [1]. As more and more cities arise, issues of resource depletion, environmental degradation, and infrastructural inefficiencies become more common [2]. In order to transcend these problems, the convergence of Artificial Intelligence (AI) and the Internet of Things (IoT) within smart city frameworks has been suggested as a solution for improving urban efficiency, resilience, or just the well-being of its citizens [3]. A smart city is considered a city that implements Information and Communication Technologies (ICT) to counter urbanization problems by building, operating, and promoting environmentally sound development practices [4]. It is essentially an intelligent network of devices and appliances that communicate wirelessly and send their data through wireless technologies and cloud computing. This wireless and cloud computing infrastructure enables real-time data gathering and analysis for planners and administrators for informed decision-making to optimize resource utilization and improve public services [5,6,7,8]. The convergence of AI and IoT technologies is being leveraged to resolve the management systems of Smart Cities. These systems provide data-driven answers to urban issues, such as reducing traffic accidents, improving energy efficiency, or detecting maintenance faults of potholes or environmental hazards. A case in point is the study done in Aveiro City, wherein these systems have exhibited their capabilities in offering actionable inputs for improving security, energy efficiency, and sustainability from the perspective of AI and IoT-based solutions for smart cities [9,10]. In addition to doing this, a study has been conducted regarding the integration of blockchain with AI and IoT, with the intention of further development of secure IoT applications. Thanks to this decentralization, essential for data integrity and security, which in turn are fundamental in the credible functioning of smart city infrastructures, blockchain guarantees a reliable environment. Its inclusion drives technological innovation in intelligent and green IoT applications through resilient architecture for urbanization [11,12]. AI and digital technologies represent both opportunities and threats to the construction of smart green units in the Global South. Qualitative research through systematic literature reviews and case studies has addressed the opportunities and challenges issued on digital technologies and AI in this regard. It probed into the policy implications and furnished a framework for smart sustainable cities alongside a call for context-sensitive approaches [13,14]. Recent research has seen frameworks emerge that aim to improve city management and active participation of citizens [15, 16]. The function of AI in intelligent city systems also encompasses other applications such as predictive analytics and infrastructure construction [17, 18]. AI can optimize energy consumption, mitigate traffic congestion, and enhance pollution management through its analysis of huge amounts of data [19]. Although there are potential advantages, the integration of IoT and AI in smart cities also poses risks, most notably data privacy and infrastructure security [20,21]. As cities grow more dependent on AI, managing these risks is important to make sure that technological progress adds value to urban development without undermining citizens' rights or public safety [22]. The research aim is the development of a Smart IoT framework combining

Information Systems and Artificial Intelligence for sustainable smart cities.

A. Motivation and Objectives

The goal is to improve the efficiency of IoT-based Smart city environments to overcome issues like minimizing congestion, security and privacy challenges, feature selection issues, uneven resource distribution, and so on. Numerous major challenges contemporaneous in existing work are labelled as follows:

- Lack of Proper Feature Selection: Existing research does not incorporate the effective feature selection approaches that lead to minimizing the model accuracy and efficiency.
- Challenges in Feature extraction: In the existing research, the robust feature extraction techniques were absent, which limits the ability to analyse and process the IoT-generated data efficiently.
- Ineffective Load Balancing and QoS Support: Existing research negatively impacts the server load balancing and fails to maintain the essential quality of service (QoS).
- Challenges in Congestion Management: Uneven distribution of IoT resources and congestion in smart city networks remain unresolved, impacting the overall system performance.
- Inadequate Threat Classification and High False Positive Rate: In the existing research, the detection mechanism lacks proper classification models, which leads to an increased false positive rate and minimizes the cybersecurity effectiveness.
- Blockchain challenges: The existing research incorporates the blockchain into IoT-based smart cities, presenting the challenges, especially in terms of high energy consumption and privacy concerns.

Motivated by the above-mentioned issues, we propose to develop a Smart IoT framework combining Information Systems and Artificial Intelligence for sustainable smart cities. This framework aims at resource optimization, feature selection, and extraction, incorporating security, improving data transmission efficiency, and ensuring load balancing and congestion avoidance for sustainable urban development. The enduring purposes of this study are labelled as follows.

- To develop an optimized feature selection approach to enhance model accuracy and efficiency in IoT-based smart city environments.
- Develop a robust feature extraction framework to enhance the data representations and analysis in IoT-based smart cities.
- Design an effective load balancing mechanism to make sure the seamless data routing and enhance the response time while maintaining the QoS.
- Develop an adaptive resource allocation approach to address the uneven distribution and minimize the congestion in IoT-based smart city networks.
- Develop energy-efficient and privacy-preserving blockchain approaches to improve the security and minimize the computational overhead in IoT-based Smart cities.
- Develop an advanced classification model for threat detection with reduced false positive rates, ensuring improved security in smart city infrastructure.

B. Research Contribution

The primary goal of this research work is the development of a Smart IoT framework combining Information Systems and Artificial Intelligence for sustainable smart cities. The following are the main contributions to this research;

- First, we utilize the CICIDS2017 dataset and then apply the normalization for preprocessing.
- After that, select the features using the wrapper technique, then apply the autoencoder for feature extraction.
- To distribute the load without any delay, we utilize the QoS-based SDN network.
- To avoid congestion when sharing the load to the server, we utilize the stateless Q-learning algorithm.
- After preserving and reducing the energy consumption in blockchain, we utilize the Hyperledger Fabric blockchain model with Linear Network Coding (LNC).
- Finally, we detect the attacks by using the Quasi-Recurrent Neural Network (QRNN) with Butterfly Optimization Algorithm (BOA) in a smart city environment.

C. Paper Organization

The subsequent components of this task have been arranged as follows. The methodologies and the literature evaluation are presented in Section II. Section III contains the problem statement and description. The proposed research method, along with the protocol and algorithm, is outlined in Section IV. In Section V, simulation results can be obtained alongside a comparison with the proposed method and other methodologies. The suggested method is depicted in Section VI.

II. Literature Review

This section summarizes and inspects the main research gaps addressed in the existing works. Authors in [20], introduce several approaches of "IoT-fog-cloud for load distribution". Second, effective fog computing techniques for load balancing amongst fog devices are provided. Last but not least, a comparison of the suggested technique's efficiency with that of current methods makes it abundantly evident that it achieves enhanced efficiency, good use of energy, excellent throughput, and efficient use of resources while lowering reaction times. However, as the volume of IoT data continues to rise at an exponential rate, effectively managing, processing, and securing this information will present significant challenges. Authors in [21] suggest a smart, resilient IoT paradigm based on AI. Each of the many services offered in the context of smart cities has begun to play a crucial part in their growth. The service takes

into account non-priority services that serve several areas, as well as route prioritization. Both services will have a highly visible location thanks to the suggested model. Furthermore, in the future, with the advent and proliferation of smart cities, there will be difficulties in achieving robust security; a cutting-edge cryptography-based model will need to be introduced to secure IoT data and infrastructure. Authors in [22] develop a next-generation "IoT-enabled Dynamic Food SCM system for smart cities" that can trace the origins of contaminants in food carrying marketplaces, intelligently route trucks, and guarantee food quality. Additionally, an IoT-based intelligent sensor-based gathering strategy is used, which is intended to increase the supply chain network's functioning effectiveness and precision. This will also increase the dataset's adaptability and achievable size, which will be accompanied by a vehicle the routing technique that tracks the causes of contaminant of contaminated food in the marketplaces. However, managing cloud computational assignments efficiently is going to be a problem. This will require the execution of assignments at the network edge to alleviate delays and conserve energy in smart city settings. Authors in [23] propose an innovative energy control and harvesting method based on hot spot concerns to improve the network lifespan of "IIoT in smart cities". The addition of energy-harvesting nodes in the standard sensor nodes has enhanced the system's "energy-aware cluster" and forwarding procedures. A thorough simulation of the suggested system is conducted under a variety of network topology scenarios. A comparison of the simulated outcomes of several of the techniques will be explored. The substantial improvement in network longevity above the current state has been demonstrated. Furthermore, challenges of load balancing are faced in smart city environments as a result of improper positioning of Energy Harvesting (EH) nodes. Seamless coverage to monitor the target area continuously became a pivotal issue. The number of Cluster Heads (CHs) or EH nodes in the present method requires, therefore, brute-force determination, which decreases scalability and requires intelligent optimization strategies. Authors in [24] suggest a structure for health monitoring and diagnostics for "geo-distributed edge clusters" that handle the large amounts of data produced by apps in smart cities. This system uses edge clusters spread throughout the smart city and is based on the MapReduce methodology for distributed large data processing. Within this architecture, a monitoring tool called SmartMonit is used to gather data on the health of devices at the edge and forecast any malfunctions using a self-organizing map technique based on artificial neural networks. The suggested approach is implemented across various clusters to evaluate its effectiveness in detecting failures. However, in the future, diagnosing failures in resource-constrained IoTs will be challenging due to the complexity of identifying root causes. Hence, the development of fault diagnosis and self-healing schemes based on root because analysis will eventually contribute to ensuring system reliability and resilience. Authors in [25] proposed to identify control over user validity in systems using a dispersed authentication method, which is considered in order to offer a safe framework for the smart computational susceptibility in the long-term sustainability of smart cities. Using a "decentralized authorization algorithm (DAA)", a "secure trust aware philosopher privacy and authentication (STAPPA)" paradigm is illustrated and put into practice. This aids in identifying and countering various smart edge computing threats that compromise data security, privacy, and the authentication of users. Additionally, during the learning phase, a "Genetic Algorithm-based Reinforcement Learning (GARL)" approach is used for optimization of networks, where it looks for abnormalities and finds the shortest path on the network site. Nevertheless, big data in smart cities is facing ballooning security threats from different kinds of attacks and vulnerabilities embedded within fog and cloud computing environments. One of the major issues at hand is to ensure the security, immutability, and decentralization of big data, thus requiring a range of lightweight authentication algorithms that can assist in the detection and mitigation of threats. Authors in [26], present a variety of approaches to create a novel safety plan for systems in smart cities. The produced "hash function, private key, public key, and session key" are all part of the "Collaborative Mutual Authentication (CMA)" method, which is utilized here to verify user identification. Furthermore, to identify network threats and guarantee the privacy of the smart city, a "meta-heuristic genetic algorithm" called Random Forest is employed. In the longer view, adoption of the suggested distributed ledger blockchain framework will be challenging for real-life applications, because it requires enormous funding and resource allocation.

III. Problem Statement

Specific existing issues, authors in [27], investigated how deep learning-based contemporary agriculture is supported by agricultural information systems in terms of the advantages of agricultural output. The agricultural IoT technology architecture served as the foundation for the suggested information service system. A deep belief neural network-based model for predicting the price of agricultural products was developed. Simulations were used to determine the DBN's ideal structure. Vegetable price prediction in a specific city was used to assess the effects of agricultural technology on the advantages of vegetable cultivation. It was found that the DBN-based agricultural product price projection system was able to estimate the price of vegetables in a given city with a high degree of accuracy. The main issues are given below:

- Their research lacks proper feature selection, which impacts the accuracy and effectiveness of the model.

Authors in [28] proposed to develop smart cities toward a more functional sharing and interconnecting, this study attempts to analyze the IoT and cloud computing-based smart city data technologies. First, IoT data pertaining to smart cities is gathered. Next, the data so gathered is normalized. For machine learning, the IoT paves the way for interaction between devices without human intervention. So, they use a machine learning algorithm known as the Adaptive Random Forest. A multi-criteria optimization is a method that they suggest for effective resource allocation of the virtual machine in Cloud Computing. They analyze and further simulate the use of IoT and Cloud Computing technologies in the development of smart cities to highlight their efficacy. The major issues are given below:

- Their research lacks feature extraction in the IoT-based Smart city environment.

Authors in [29] proposed to improve communication efficiency and dependability with the least amount of delay in terms of energy use and data privacy for data transfer. This study presents "Intelligent Bison-based secure edge-enabled computing IB-SEC)" architecture for smart cities leveraging the IoT. To increase the efficiency, security, and dependability of data transfer in an Internet of Things smart city network, the created IB-SEC platform combines a distributed hashing-based security method with the "African Buffalo Optimization" method. To be able to provide a secure wireless connection and safeguard data from unwanted access, this platform integrates robust edge computing and "MAC (median access control)" protocols for identification, authentication, encryption, and access control. In summary, the IB-SEC framework offers a cutting-edge strategy for safe Internet of Things networks and allows smart city applications. Additionally, the platform that has been designed enables the integration of diverse "edge devices, sensors, and systems" that are efficiently managed and adjusted by MAC protocols. The main issues are given below:

- Their research negatively impacts the load balancing on the server and does not support the quality of services such as data transmission and response time.

Authors in [30] suggested the behaviour of SUs in the context of poor sensing is mathematically modelled. In their suggested DNN-based CR-IoT designs, they presented a complete spectrum preference methodology based on DNN. For an intelligent choice, they have also suggested a novel DNN-based complete spectrum decision method to meet the spectrum decision framework. This allows one to choose, depending on the quality of the accessible channel, whether to proceed with the data communication or to forego the chance to and wait for the subsequent sensing event. When there is a large density of IoT devices in a relatively tiny area with an elevated need for spectrum, their study will prove extremely helpful. Here, the primary challenges are given below:

- Their research fails to effectively address the challenges in the uneven distribution of resources and congestion in the IoT-based smart city network.

Authors in [31] proposed the OMLIDS-PBIoT approach, a unique optimum "Machine Learning-based intrusion detection system" for secure BioIoT in a smart city setting, which has been created in this paper. Data pre-processing is carried out in the subsequent step to convert data into the proper format, and the OMLIDS-PBIoT approach functions by employing BC and ML techniques. Additionally, an FS model that utilizes "golden eagle optimization (GEO)" is created in order to extract relevant subsets. For intrusion categorization, the "random vector functional-link network (RVFL)" and HBO models were used. In turn, the use of blockchain is used for safe data transfer in the context of IoT-enabled smart cities. Here, the primary challenges are given below:

- Their research lacks proper classification of threat detection and fails to minimize the false positive rate.
- Their research faces challenges in blockchain implementation, especially concerning energy consumption and privacy issues.

Recent studies have further expanded the scope of AI-enabled sustainable urban systems. Faisal et al. [32] integrated machine learning with multi-criteria decision making to assess AI-driven personalization in smart cities. Huy and Phuc [33] examined how technologically vigilant leadership supports smart, sustainable circular supply chain management in Industry 6.0 environments. Xiao and Dong [34] optimized AIGC technology for IoT devices using deep learning, highlighting efficient AI deployment on resource-constrained urban edge nodes.

A. Research Solution

The Wrapper technique is used for feature selection and autoencoder for feature extraction in IoT-based smart city environments. A QoS architecture based on SDN balances server load distribution and manages response time and data transmission. Stateless Q-learning avoids congestion. Hyperledger Fabric blockchain preserves privacy and reduces energy consumption. Quasi-Recurrent Neural Network and butterfly optimization algorithm are combined for attack classification and accuracy in IoT-based smart city environments.

IV. Proposed Work

The primary goal of this research work is the development of a Smart IoT framework combining Information Systems and Artificial Intelligence for sustainable smart cities. Fig. 1 represents the overall architecture of the proposed methodology. As part of this strategy, the following key steps are taken

- Information Gathering using IoT devices
- Feature Selection and Feature Extraction
- Quality of service-based load distribution
- Congestion avoidance in a smart city environment
- Privacy-preserving, energy consumption-based smart city using blockchain
- Artificial Intelligence-based attack detection in smart cities

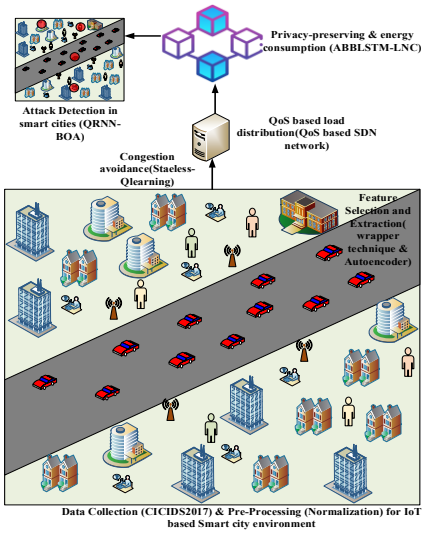


Fig. 1. The overall architecture of the proposed methodology.

The end-to-end workflow of the proposed framework operates in two coordinated evaluation paths. In the intrusion-detection path, CICIDS2017 data are normalized, filtered through wrapper-based feature selection, encoded by the autoencoder, and classified by the BOA-optimized QRNN to detect cyber-attacks. In the network-simulation path, NS-3.35 generates smart-city IoT traffic across SDN-enabled servers, where the controller performs QoS-based load balancing and stateless Q-learning reduces congestion. Detected security events and validated network transactions are recorded in Hyperledger Fabric, while Linear Network Coding reduces blockchain transmission overhead and energy consumption. This separation ensures that each evaluation component is supported by an appropriate data source and simulation environment.

A. Information Gathering Using IoT

Initially, we collect the dataset for our framework. We utilize the CICIDS2017 dataset. Although CICIDS2017 is an intrusion-detection benchmark derived from controlled network traffic rather than direct smart-city IoT operational traces, it is used in this study to evaluate the cyber-attack detection component of the proposed framework. Network performance metrics, including response time, throughput, congestion control, and energy consumption, are evaluated separately through NS-3 simulations, as described in Section V. Before further processing, we implement the normalization approach to standardize data values, make sure that variations in scale do not affect the model performance. By employing this dataset with efficient preprocessing, we can design the real-world scenario and improve smart city resilience against potential threats. Normalization

Normalization is a preprocessing step to scale input data so model has ranges consistent (increased performance). A standard formula for normalization is:

$$M_{norm} = \frac{M - M_{min}}{M_{max} - M_{min}} \quad (1)$$

where,

M is the initial feature value,

M_{min} is the minimum value of all of the feature,

M_{max} is the maximum value of the feature, and

M_{norm} is the normalized feature value.

This normalizes all feature values between 0 and 1 ensuring that higher magnitude features.

B. Feature Selection and Feature Extraction

With an extensive dataset, the next crucial step is identifying the most significant features. So, we utilize the Wrapper based feature selection approach, we filter out redundant and irrelevant data, ensuring that only the most important attributes are utilized for analysis. To further refine the process, we apply the Autoencoder for feature extraction, enhancing the model performance and achieving the high accuracy. This step not only improves the predictive capabilities but also enhance the overall effectiveness in decision-making within the smart city environment.

1) Wrapper-based feature selection: Wrapper-based feature selection is an iterative algorithm that analyzes sets of features by training a machine learning model and then assessing its accuracy. The method reconciles accuracy with computational power by picking the most relevant features and eliminating redundant and irrelevant information. The CORRAccuracy algorithm is an example of a highly acclaimed wrapper-based approach that uses correlation-based filtering and accuracy assessment to pick optimal features.

The CORRAccuracy measure estimates the quality of features selected based on the following formula:

$$CORR = \frac{V \times avg(corr_{hc})}{\sqrt{V + V(V-1) \times avg(corr_{hh})}} \quad (2)$$

where,

$corr_{hc}$ is the feature-class correlation,

$corr_{hh}$ is the correlation between features, and

V is the number of features selected.

This measure ensures that the chosen features have high correlation with the target class while reducing inter-feature redundancy. After correlation is calculated, the accuracy measure (*accuracy*) also confirms feature importance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Here,

TP denoted as true positive

TN denoted as true negative

FP is denoted as false positive

FN is denoted as false negative

The wrapper method is one where training and evaluation are repeated cycles in order to locate the optimal subset of features that maximizes classification performance. Yet another critical feature selection equation is the information gain (IG), which measures the extent to which a feature reduces uncertainty:

$$IG = I(A) - I(A|B) \quad (4)$$

where,

- $I(A)$ Represents the target variable of entropy
- $I(A|B)$ Denotes the entropy of the target given the feature B

Mutual information (MI) is also used for ranking features by quantifying the dependence between a feature and the class:

$$MI(A,B) = \sum_a \sum_b JP(a,b) \log\left(\frac{JP(a,b)}{JP(a)JP(b)}\right) \quad (5)$$

Here,

- $MI(A,B)$: Mutual Information between feature A and target class B.
- $\sum_a \sum_b$: Summation over all possible values of a and b.

- *JP*(*a*,*b*): Joint probability distribution of feature A and class B (i.e., the probability that feature A takes value a and class B takes value b at the same time).
- *JP*(*a*): Marginal probability of feature A (i.e., the probability that A=a happens).
- *JP*(*b*): Marginal probability of class B (i.e., the probability that B=b happens).
- *a*: Denotes an instance of the feature A.
- *b*: Denotes an instance of the class B.

In order to optimize the feature subset, the feature ranking function combines correlation and accuracy as:

$$\text{Rank}(l) = \beta \times \text{CORR}(l, \text{class}) + \emptyset \times \text{Accuracy}(l) \quad (6)$$

- *Rank(l)*: Final ranking score of feature *l*.
- *l*: Individual feature being examined.
- β: Correlation metric weighting factor (governs the relative importance of correlation in ranking).
- *CORR*(*l*, *class*): Feature *l* correlation with the target class.
- $\emptyset$: Accuracy metric weighting factor (governs the relative importance of accuracy in ranking).
- *Accuracy(l)*: Model accuracy when the feature *l* is used in the classifier.

This formula orders features by blending correlation with the target class and with the classification accuracy obtained using the feature. The weights β and $\emptyset$ enable the fine-tuning of the impact of each measure in the ordering. For $\beta > \emptyset$ correlation is given greater precedence, and $\emptyset < \beta$ gives a greater precedence to accuracy. This strikes a balance between choosing features that are informative as well as beneficial towards model performance.

2) Autoencoder for feature extraction: A unique kind of multilayer perceptron called an autoencoder has the same number of neurons in both its input as well as its output layers. Two elements of the autoencoder's design, the encoder and the decoder, are taught at every succeeding layer. In a deep autoencoder, each layer receives input from the layer before it. Fig. 2 represents the autoencoder architecture. In particular, the autoencoder has been taught to convert the raw input into a compressed representation, from which the output is rebuilt.

$$I(y) = \varphi(\mathbb{C} \cdot z(y) + c) \quad (7)$$

where,

$I(y)$ is the encoded of the input y, $z(y)$ is the input data, φ is the activation function, the acquisition function in this instance is sigmoid. The letters "weight" and "bias" are $\mathbb{C}$ and c, correspondingly.

$$S(y) = b(\mathbb{C} \cdot \text{H}(y) + c') \quad (8)$$

H(y) is denotes latent representation produced by the encoder.

c' is denoted as bias implemented during decoding.

b(.) is used to ensure the output stays within a range.

$S(y)$ is the predicted output of the input $I(y)$ based on the latent representation H(y). The weights of the autoencoders are optimized to reduce the inaccurate reconstruction of the network. The reconstructed error is calculated as follows:

$$(z(y), S(y)) = \|\text{E}(y) - S(y)\|^2 \quad (9)$$

Here, E(y) is the actual input data. $\sum_{E=1}$. It represents the summation and the index starts with first element. Furthermore, the cross-entropy measure's reconstruction error for binary values (7):

$$\text{h}(z(y), S(y)) = -\sum_{E=1} \left[z(y)_E \log S(y)_E + \left(1/(y)_E\right) \log\left(1 - S(y)_{s_E}\right) \right] \quad (10)$$

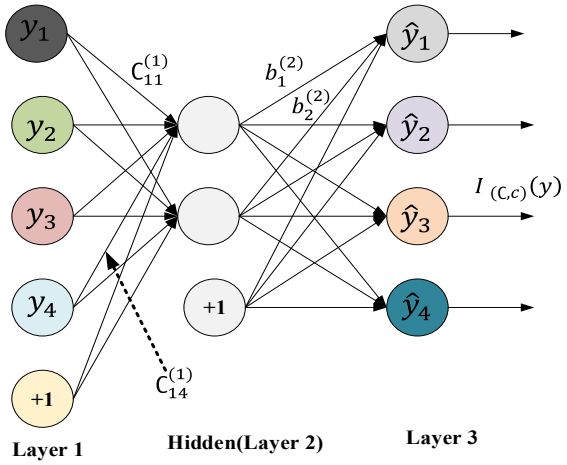


Fig. 2. Autoencoder.

C. Quality of Service-Based Load Distribution

Based on the selected and extracted features, the system processes only the most relevant data, minimizing computational overhead and enhancing the response times. However, smart city generates an enormous volume of data, leading a probable server congestion and latency. To address this, we employ the Quality of Service (QoS) mechanism using software defined networking (SDN). This technique ensures the dynamic load balancing, optimizing server response time and effective data transmission. By distributing the loads, our framework mitigates the bottlenecks, enabling real-time decision making across the city's interconnected system.

SDN was selected over conventional distributed routing and static load-balancing approaches because it provides centralized visibility of network state, policy-based QoS enforcement, and programmable traffic engineering in large-scale IoT deployments. Compared with traditional IP-based management, SDN enables faster adaptation to congestion, more efficient resource allocation across servers, and lower end-to-end latency under dynamic smart-city traffic conditions.

Quality of Service (QoS) mechanism using software-defined networking (SDN).

In the Software Defined Networking (SDN) based Quality of Service (QoS) mechanism for load balancing, the main objective is to dynamically allocate incoming requests to multiple servers so that server response times are optimized and data transmission is efficient. Since smart city systems produce a vast amount of data, the SDN architecture is responsible for preventing server overload and reducing latency. In order to realize dynamic load balancing, the SDN controller is always observing the workload of every available server and sends the incoming requests to the server that has the lightest workload. The following describes the load distribution with a mathematical equation:

$$R_o = \min(R_1, R_2, \ldots, R_p) \tag{11}$$

where,

R_o is the load for the server selected and $R_1, R_2, \ldots, R_p$ describes the present loads on all the available servers. The formula keeps, in a dynamic manner, forwarding new requests to the server with the least load and thereby avoids overloading any single server. By constantly evaluating the load, the system developed in SDN efficiently spreads traffic across all available resources, ensuring no bottlenecks are created.

In addition, there is another load threshold condition to further enable optimization of load distribution. Each of the servers has been given a maximum load past which it cannot be considered efficient; when the load on that server exceeds this threshold, requests are diverted to other servers that are less loaded. This is what is mathematically expressed as:

$$R_o \leq R_{max} \tag{12}$$

R_o load the server o while R_{max} is an allowable maximum load. It stops any single server from being saturated or loaded, giving the system the ability to dynamically reroute any excess requests forwarded to it, thus optimizing request processing and availability of continuous service.

A key performance measure in load distribution under QoS is response time Res_{time}, defined as the time elapsed from the receipt of a request until a response is issued. It is calculated with the formula:

$$Res_{time} = time_{end} - time_{start} \tag{13}$$

where, $time_{start}$ represents the time at which the request is received, and $time_{end}$ corresponds to the moment when the system finishes processing and returns. Reducing response time is critical in ensuring real-time decision-making capacity, particularly in smart city functions where there is a need for quick processing of activities like traffic monitoring and emergency response. Through the dynamic allocation of requests to idle servers, the SDN controller minimizes delays and maximizes the capacity of the system to respond quickly to arriving data.

Another key system performance metric is server utilization, which measures the percentage of a server's processing power being utilized. This is determined by the formula:

$$U_o = \frac{time_{busy}}{time_{total}} \times 100 \tag{14}$$

where U_o represents the utilization percentage of the server o, $time_{busy}$ is the time the server spends actively processing requests, and $time_{total}$ is the overall observation time. The measure helps in verifying the resource optimization of the servers. To distribute the workload among several servers, the SDN controller prevents overloading of any server and keeps most of the running states steady and effective. The method also helps in synonyms scaling the system to accommodate the increasing data load without any performance degradation.

An additional load difference function runs within the SDN controller to maintain the load balance over multiple servers. This can be represented mathematically as:

$$\Delta Y = \left| Y_q - Y_o \right| \tag{15}$$

where, ΔY is the difference in load between the server q and server o. Whenever that difference exceeds a pre-set limit, the SDN controller dynamically allocates incoming requests from the overwrought server into the less-loaded server place. Such load management in advance avoids congestion, increases server performance, and guarantees that the entire system keeps its responsiveness uniform even when traffic fluctuates over a certain limit.

Further, the system analyzes its performance based on the data rate of transmission, i.e., the rate at which data is transferred successfully within a time limit. It is determined through the formula:

$$D = \frac{S_{transmitted}}{time} \tag{16}$$

The data transmission rate is expressed as D with $S_{transmitted}$ being the number of packets that have been successfully transmitted, and time being the observation time of data transmission. The faster the data transmission rate translates to a larger amount of data processed effectively by the system, thereby enhancing the quality of service for smart city infrastructures. Through constant checking and fine-tuning of the data transmission rate, the SDN controller encourages rapid and timely transmission of data and supports real-time data applications in surveillance, traffic flow, and public safety.

The QoS-based load distribution scheme based on SDN efficiently meets the demands presented by the mass data generation of smart city applications. With the deployment of dynamic load balancing equations, such as the minimum load selection, load threshold condition, response time computation, server utilization tracking, load difference function, and data transmission rate, the system provides enhanced server performance, less latency, and effective resource management. This integrated approach allows the SDN framework to have high responsiveness and reliability levels, supporting real-time decisions and improving the general level of service quality in smart city interconnecting systems.

D. Congestion Avoidance in Smart City Environment

The uneven distribution of resources leads to congestion and unhitches the whole system of smart city operations. To combat this congestion, we will deploy stateless Q-learning, a reinforcement learning-based approach that will dynamically adjust the allocation of computational resources. Thus, this becomes a decentralized technique that will involve building up

a congestion avoidance methodology that will work in the shadows to prevent any conceivable congestion from disrupting IoT services like traffic management, energy distribution, and emergency response from functioning smoothly.

1) Stateless Q-learning: Stateless Q-learning is a reinforcement learning technique that can be applied in situations where it is not practical to store a complete representation of the system state. In an IoT smart city scenario, with millions of IoT devices constantly producing computational loads, stateless Q-learning is perfect for real-time congestion avoidance. It enables IoT devices to make autonomous adjustments to their computational resource allocation without needing a central controller, lessening communication overhead and ensuring real-time response.

The Q-learning here operates by modifying a Q-value—a value that estimates the expected future reward gained by performing an action. For every IoT device δ at time step, the Q-value will be calculated by the following formula:

$$Q_i^{\delta}(time)=\big(1-X_{\delta}(time)\big)Q_i^{\delta}(\tau\text{-}1)+X_{\delta}(time)l_i^{\delta}(time) \quad if\ i=U_{\delta}(time) \tag{17}$$

Otherwise,

$$Q_i^{\delta}(time)=Q_i^{\delta}(time\text{-}1) \tag{18}$$

Here,

$Q_i^{\delta}(time)$ is the current Q-value for action (i) by device δ at time.

$X_{\delta}(time)$ is the learning rate, which determines how much of the new information replaces the old Q-value.

$U_{\delta}(time)$ is the action selected by device δ at time *time.*

$l_i^{\delta}(time)$ is the reward received by device δ for taking action i.

Reward function to optimize for completing tasks in a minimum amount of time. At a device δ, reward at a time is determined by:

$$l_i^{\delta}(time)= TH_{\delta}\text{-}T_{\delta}(time) \tag{19}$$

Here,

TH_{δ} is the threshold of time by which the task has to be achieved.

$T_{\delta}(time)$ is the real-time taken by the device to get the task accomplished.

A positive reward is issued if the task is achieved within the threshold time, and a negative reward is issued if the task takes more time than the designated time. It encourages the system to choose such actions that bring down congestion as well as lessen latency.

The agent chooses actions based on an ε-greedy policy to trade off exploration and exploitation. With probability $1\text{-}E_{time}$, the best action according to Q-value is chosen, and with probability E_{time}, a random action is sampled:

$$U_{\delta}(time)=\begin{cases}\arg\max Q_i^{\delta}(time\text{-}1) \ with\ probability\ \ 1\text{-}E_{time}\\ random\ action\ \ with\ \ probability\ E_{time}\end{cases} \tag{20}$$

where,

$$E_{time}= E_0 \times time^{-1/\delta} \tag{21}$$

Here, E_{time} reduces as the algorithm continues, enabling early exploration and subsequent exploitation of the best-known actions.

The learning rate $X_{\delta}(time)$ is given as:

$$X_{\delta}(time)= \left[\mu+ \Delta_{\tau}^{\delta}\big(U_{\delta}(time)\big)\right]^{-\omega} \tag{22}$$

where,

μ is a constant regulating the initial learning rate.

Δ_{time}^{δ} is the number of times action

$U_{\delta}(time)$ has been chosen up to time *time.*

ω is a parameter regulating the rate at which learning diminishes.

As the learning rate decays over time, the policy gradually stabilizes and selects actions that reduce congestion under the current network conditions. Decentralized decision-making is facilitated by this that IoT devices optimize resource usage autonomously, minimizing congestion and maximizing the system's efficiency overall. With the application of stateless Q-learning, intelligent city infrastructure can dynamically distribute computation resources on a network of IoT devices to provide low-latency task execution and prevent service failure in high-priority domains such as traffic control and emergency response. Fig. 3 represents the Q-learning.

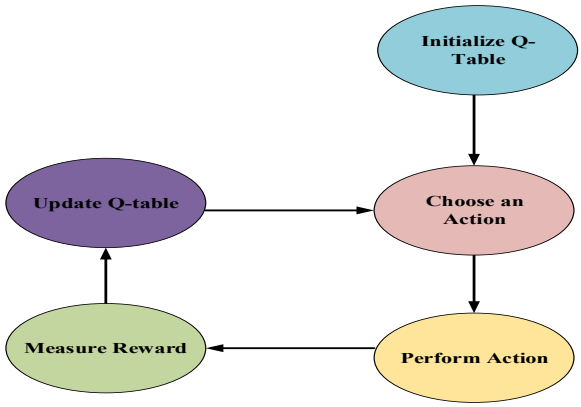


Fig. 3. Q learning.

E. Privacy Preserving-Energy Consumption-Based Smart City Using Blockchain

Blockchain technology promotes heightened security in smart city ecosystems, and it also introduces added challenges pertaining to energy consumption and privacy preservation. Such problems can be addressed by merging the Hyperledger Fabric blockchain with Linear Network Coding (LNC). HLFB ensures robust privacy protection of transactional data analysis and security, while Linear Network Coding improves on energy-

efficient transmission. The combination of both enhances security without compromising sustainable blockchain smart cities.

1) Hyperledger fabric blockchain: In the Hyperledger Fabric blockchain system, an efficient privacy-preserving mechanism is created through several cryptographic methods and operational stages. In the initialization stage, key generation and authentication security is taken care of by the Membership Service Provider (MSP) using random numbers and a hash function. The process of registration entails creating and authenticating unique passwords and identifiers through clients, smart contracts, and endorsing peers. For example, in the process of smart contract registration, a verification message vf_1 is generated using the following equation:

$$vf_1 = \left(e\left(r_x^* F\right) \| F\right) \oplus \mathrm{h} \tag{23}$$

Here, $r_x^* F$ represents the hashed smart contract ID; F is any random value, while h is the verification parameter. Now, this message will be kept by the MSP and committing peers for future verification.

A similar verification message vf_2 is generated at the time of client registration:

$$vf_2 = \left(e\left(Ct_x^*\right) \| rv\right) \oplus avf \tag{24}$$

where, Ct_x^* the client's is hashed ID, rv is a random value, and avf is the related verification parameter. This maintains strict verification procedures for each entity that registers to prevent any unauthorized access. Encryption of data protection is carried out using the Advanced Encryption Standard (AES) technique, which has a key size of 128 bits, where E_i would be the encrypted information as follows:

$$E_i = E(rd, ek) \tag{25}$$

where,

rd is record while ek is encrypted key. For maintaining better privacy, the key ek itself is generated using an interpolation-based approach:

$$ek = YI \oplus (C * F) \tag{26}$$

Here, YI is an interpolated value that is computed as:

$$YI = YI_1 + \frac{(YI_2 - YI_1)}{(XI_2 - XI_1)} * (XI - XI_1) \tag{27}$$

XI refers to the input value that you are finding the interpolated output for.

XI : The lower boundary (beginning point on the X-axis).

XI_1: The upper boundary (terminal point on the X-axis).

C is computed as

$$C = 60X^3 + 6X^2 + 6 \tag{28}$$

This way advanced calculations ascertain that the encryption keys are cryptographically derived with the uniqueness of being different from one-another and the keys cannot be accessed by third-party entities externally. Also, the authentication process involves generating and validating digital signatures:

$$sg = (e(mn_x F) \| j) \bmod k \tag{29}$$

sg is the digital signature.

$mn_x F$ is message data augmented with a random value or system-defined parameter

$e(\cdot)$ is the application of a hashing or encoding function to secure the input.

$\|\ \|$ is concatenation of the encoded message with some other value

j (which can be a random nonce or session-specific identifier).

$mod\ k$ makes sure that the signature is computed in a modular arithmetic framework, which is usually employed to restrict the length of the signature and increase cryptographic security.

This guarantees the integrity of the participants in a blend of hashing and modular arithmetic. The authenticity of every message of authentication is guaranteed by means of time-sensitive verification, providing yet another security layer to the process.

In total, the Hyperledger Fabric blockchain platform employs these strict mathematical operations to provide safe registration, data encryption, and verification, facilitating strong privacy protection in smart city systems.

2) Linear network coding: In linear coding, the source node divides the data into several blocks. There are n packets, referred to as native packets, in each block. The symbol for this is z_k, Z{1,2,…,n}. As a result, the source node sends the LNC-coded packet $z_j^{'}$ to the subsequent CH. The mathematical representation of this packet $z_j^{'}$ is provided by Eq. (30).

$$z_j^{'} = \sum_{k=1}^{n} D_{jk} y_k \tag{30}$$

where,

D_{jk} is matrix of coefficients

z_k represent k th native packet.

Z{1,2,…,n} is denotes that Z belongs to the set {1,2,…,n}

n denotes as total number of packets in the block.

A Galois Field is used to perform operations such addition as well as multiplication. Code vector, $\vec{D}_j = \left(D_{j1}, D_{j2}, \ldots, D_{jn}\right)$ and the block ID are contained in the data packet's $z_j^{'}$ header. The starting value of $n^{'}$ is stored in a counter, and its value is bigger than n. The source node keeps track of this counter. The counter is reduced by one unit for every coded packet that is passed on to the next node. The random distribution of the coded data packets goes on until the value of the counter reaches up to Zero. The recruiting and forwarding operations are performed at every single hop. The destination CH initiates the formation of the next cluster upon receiving the packets of the same block. Now, this receiving cluster is converted into a new transmitting cluster that

contains the same nodes. The transmitting CH first schedules the corresponding time at which the node should transmit the encoded data to the receiver cluster, and it is utilized by each node of the transmitting cluster. The receiving CH then captures control of the surrounding nodes to constitute the receiver cluster, and to select the nodes with the higher cost D_j. This D_j variable is specified for a node j, which may possibly be used for the receiver cluster, as described in Eq. (31).

$$D_j = \sum I_{jk}\left(1-Q_{jk}\right) \forall \text{ node } j \varepsilon \text{ sending cluster} \quad (31)$$

where,

Q_{jk} – Probability of loss

I_{jk} - The indication's function is to show if the sensor is available to receive the packets.

Since it is reserved to receive or transmit data packets for another cluster from a different route, these may not be available. Every neighbor's ping sample is used to periodically test the probability Q_{jk}. In order to send the data, I_{jk} depends on the CH transmission method, which designates the sensor nodes. As a result, the transmitting cluster selects the sensors that are anticipated to receive more data as a receiver cluster.

Assume for the moment that a node has acquired the encoded data z. The newly encoded data packet can then be retrieved using Eq. (32) as well as Eq. (33).

$$z'' = \sum_{j=1}^{n} D_j z_j' \quad (32)$$

where,

D_j is arbitrarily generated numbers

z'' is the native data packets

$$z'' = \sum_{j=1}^{n} D_j \left(\sum_{j=1}^{n} D_{jk} y_k\right) = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} D_j D_{jk} y_k\right) z_j = \sum_{k=1}^{n} g_k z_k \quad (33)$$

When sending the encoded data packet, a sensor will include a new code vector, represented as $\vec{g} = (g1, g2, \ldots, g_n)$, within the packet's z'' header, much like the transmitting node does. The system will first determine if the data is new when it receives encoded data. Only when a data set is linearly independent of previously received data from the same block that the sink node has received is it deemed new and fresh. The data will be immediately deleted if it is not new. Thus, the receiver node can retrieve transformative data packets as long as fresh packets are collected. Decoding at the sink node entails applying the Gaussian technique to solve the group of linear equations. If the rank of the matrix is known, the linear equation specifies the solution technique.

$$\begin{pmatrix} D_{11} D_{21} \cdots D1_n \\ D_{21} D_{22} \cdots D2_n \\ \cdots \cdots \cdots \cdots \\ D_{n1} D_{n2} \cdots D_{nn} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{pmatrix} = \begin{pmatrix} z_1' \\ z_2' \\ \cdots \\ z_n' \end{pmatrix} \quad (34)$$

where,

z_1' is an encrypted packet

$(D_{k1}, D_{k2}, \ldots, D_{kn})$ is the code vector

z_k denotes native packets

The sensors in the sending cluster may have received two or more packets from the previous cluster, allowing them to create their own combination. Using randomly selected coefficients, the sensor nodes combine all received packets from the same block to produce a random linear combination.

F. Artificial Intelligence-Based Attack Detection in Smart Cities

Cyber security is among the major priorities of all smart city strategies. In order to enable effective detection and classification of cyber threats with minimal false positives and maximization of attack detection accuracy, we apply a Quasi-Recurrent Neural Network (QRNN). To boost performance, the Butterfly Optimization Algorithm (BOA) is integrated, which results in an accelerated learning process and higher classification speed. Thus, the synergy between QRNN and BOA provides an extremely responsive and intelligent defense mechanism against new cyber threats for smart city networks.

1) Quasi-Recurrent Neural Network (QRNN): The Quasi-Recurrent Neural Network (QRNN) is a hybrid deep model that exploits the fast feature extraction of convolutional layers and the sequential modeling capacity of recurrent networks. The architecture lends itself especially well to smart city application areas where attack detection must be computationally efficient and precise. By providing speed in computation while largely decoupling the computations across time steps, QRNN achieves enhanced performance without leaving the ability to handle sequential data.

QRNN was preferred over conventional CNN, LSTM, and Random Forest classifiers because it offers a better balance among detection accuracy, inference latency, and deployment cost in high-volume IoT traffic. Compared with LSTM-based models, QRNN reduces sequential computation overhead while preserving temporal dependency modeling. Compared with standard CNN and tree-based methods, it achieves stronger attack classification performance on sequential network-flow data with lower false positive rates, making it more suitable for real-time smart-city security monitoring.

The architecture consists of three main components: a convolutional layer that extracts features, a gating mechanism that regulates the flow of information, and a recurrent pooling layer that serves to maintain sequential dependencies

QRNN computes the victim of an attack against a data set, which entails feeding in an input sequence $P = [P_1, P_2, \ldots, P_{time}]$ where each P_t is an input feature at the time step. The convolutional layer convolves many filters over this sequence to derive spatial features. The output of the convolutional layer is given by:

$$H_t = \sigma\left(k * P_{time} + c\right) \quad (35)$$

Here, H_{time} is the feature extracted at time, k is a convolutional filter, c is the bias and $*$ refers to the operation of convolution. The activation function σ (usually sigmoid) provides the non-linearity and assists with detecting complex

patterns. The mechanism of the gate also improves upon information flow by deciding which feature should be given to the following layer. The gate value is calculated as:

$$v_{time}=\sigma\ (k_v * P_{time}+ c_v) \tag{36}$$

Here, v_{time} represents the gate value, k_v is the filter for the gating function, and b_v is the bias. This mechanism serves to control the information passed from time steps, keeping out unnecessary or noisy data from passing through the network. The recurrent pooling operation aggregates the current input and the previous hidden state, enabling QRNN to keep memory along sequences. This is expressed by the following equation:

$$g_{time}= F_{time}\odot C_{time-1}+ (F_{time})\odot H_{time} \tag{37}$$

In this equation, C_{time} represents the current memory cell at time, F_{time} is the forget gate that determines the amount of previous memory C_{time-1} that is preserved, and H_{time} is the output of new features. The notation $\odot$ is used for element-wise multiplication. Finally, the results of the LSTM unit are also acquired through a gate controlling the visibility of their internal memory to layers outside:

$$V_{time}=O_{time}\odot C_{time} \tag{38}$$

where, V_{time} is the LSTM output, O_{time} being the output gate, which decides to what extent internal memory is exposed: it is the output that is passed on for doing attack detection, where the LSTM decides whether an incoming data stream is that of normal operation or a potential instance of cyber-attack.

2) Butterfly Optimization Algorithm (BOA): In the improvement of the efficiency of the QRNN, Butterfly Optimization Algorithm (BOA) is employed for optimizing its hyper parameters. These habits by which the butterflies search for food and mate are really nature-inspired, coupled with movements driven by a sensory modality known as fragrance. The algorithm maintains a balance between global exploration and local exploitation in order to enhance the accuracy of the model as well as avoid false positives. The fragrance that signifies the objective function or the fitness is calculated as follows:

$$K= cI^A \tag{39}$$

In this equation, K is the sensed fragrance, c is the modality of the senses (a constant that determines sensitivity), I is the intensity of the stimulus, and A is the power exponent that determines the rate of absorption. The flight of a butterfly during the global search phase, when it moves towards the optimum in the population, is described by:

$$\zeta_i^{\varsigma+1}=\zeta_i^{\varsigma}+ \left(\psi^2\times J^*-\zeta_i^{\varsigma}\right)\times K_i \tag{40}$$

Here, $\zeta_i^{\varsigma+1}$ is the updated position of the ith butterfly in iteration $\varsigma+1$, J^* represents the best solution found so far, and ψ is a random number between 0 and 1 to ensure stochastic search. This equation facilitates global searching of the best hyperparameters by modifying the weights and biases of the QRNN. For the local search step, in which the algorithm searches in the vicinity of solutions, the position of the butterfly is updated with:

$$\zeta_i^{\varsigma+1}=\zeta_i^{\varsigma}+ \left(\psi^2\times\zeta_j^{\varsigma}-\zeta_k^{\varsigma}\right)\times K_i \tag{41}$$

In this formula, ζ_j^{ς} and ζ_k^{ς} are two random butterflies selected from the population. This local search improves the ability of the algorithm to fine-tune the model's performance by searching for close-to-optimal solutions. The algorithm assesses the fitness of every butterfly (or solution) based on the classification error rate:

$$\text{fitness}(\zeta)=\frac{\text{number of misclassified samples}}{\text{Total samples}}\times 100 \tag{42}$$

This task estimates the goodness of a specific solution, in which a lower error rate is reflective of improved model performance. Iteratively, BOA reduces this error, hence achieving increased accuracy and improved generalization to unobserved data.

The combination of QRNN with BOA produces a robust and smart cyber-attack detection system for smart cities. The QRNN processes big-scale sequential data effectively while keeping computational efficiency, and the BOA optimizes the performance of the model by dynamically tuning hyperparameters. This integrated system gives a strong, adaptive defense system against new cyber threats, which keeps smart city networks resilient and secure. Algorithm 1 describes the pseudocode for the butterfly optimization algorithm.

Algorithm 1: Butterfly Optimization Algorithm (BOA)

Input: N (population size), Max_iter (max iterations), c (sensitivity), A (absorption rate), QRNN (model), X, Y (data)

Initialize butterfly population (Z) with random QRNN hyper parameters

Evaluate fitness

fitness(ζ) = (Misclassified samples / Total samples) × 100

Set J_{best} = best solution with minimum fitness

For ς=1 to Max_iter:

For each butterfly ς_i:

$K= cI^A$ # calculate fragrance

If random(0, 1) < p: $\zeta_i^{\varsigma+1}=\zeta_i^{\varsigma}+ \left(\psi^2\times J^*-\zeta_i^{\varsigma}\right)\times K_i$ # Global search

Else:

Select ζ_j^{ς} a ζ_k^{ς} from population,

$\zeta_i^{\varsigma+1}=\zeta_i^{\varsigma}+ \left(\psi^2\times\zeta_j^{\varsigma}-\zeta_k^{\varsigma}\right)\times K_i$ # Local search

Update J_best if a better solution is found

Return J_best (optimized QRNN hyperparameters) and fitness(J_best)

V. Experimental Results

This section presents the performance evaluation and experimental evaluation of the suggested study strategy. This part is divided into three subsections: research summary, comparative analysis, and simulation study.

A. Simulation Study

The proposed study approach is replicated using NS 3.35 with Python. This tool works well and offers all the specifications needed for the recommended approach. In Tables I and II, the system specification and parameters are presented.

TABLE I. SYSTEM SPECIFICATIONS

Software Specifications	OS	Ubuntu 22.04
	Network Simulator	NS 3.35 with Python
Hardware Specifications	RAM	4GB
	Hard Disk	500GB

TABLE II. SIMULATION PARAMETER

Parameters		Descriptions
Network Parameters	IoT Devices	50
	Base station	2
	Blockchain node	1
	Server	3

B. Intrusion Detection Experimental Setup

The QRNN-based attack detection experiments were conducted on the CICIDS2017 dataset using an 80/20 stratified train-test split. After preprocessing and normalization, 500,000 flow records were used for training and evaluation across 14 attack classes. Wrapper-based feature selection retained 45 features, which were encoded by an autoencoder with a 32-dimensional latent representation. The QRNN consisted of two convolutional layers (128 and 64 filters), a quasi-recurrent pooling layer, and a dense output layer. BOA optimized the learning rate (0.0001–0.01), batch size (16–128), and number of epochs (10–50) over 30 iterations with a population size of 20. The final model was trained with a batch size of 64, a learning rate of 0.001, and 50 epochs. Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and a confusion matrix.

C. Comparative Analysis

This section relates the recommended approach to several current ones, such as Genetic Algorithm-Based Reinforcement Learning (GARL) [21], IoT–fog–cloud (IFC) [16], Intelligent Buffalo-based Secure Edge-enabled Computing (IB-SEEC) [25], Advanced Random Forest (ARF) [24] and the proposed approach measures its efficiency using performance metrics such as Response Time (ms), Throughput (Mbps), Attack detection (%), False Positive rate and Energy Consumption (%).

For the network simulation experiments, the baseline methods were implemented in NS-3.35 under the same topology, traffic model, and hardware settings summarized in Tables I and II. For the intrusion-detection experiments, all methods were trained and tested on the same CICIDS2017 partitions using identical preprocessing, feature selection, and evaluation metrics.

1) IoT devices load (Mbps) vs. response time (ms): The correlation between Response Time (ms) and Total Load (Mbps) in an intelligent IoT environment can be described as:

$$RTS = \frac{TL}{AB\text{-}Uz} \tag{43}$$

where,

RTS = Response Time (ms).

TL = Total Load (Mbps) (data arrival rate).

AB = Available Bandwidth (Mbps).

Uz = Current System Utilization (Mbps).

With higher Total Load (TL) or reduced Available Bandwidth (AB), the Response Time RTS grows with congestion and processing delays. Keeping Uz low ensures quicker response times by avoiding system saturation.

TABLE III. NUMERICAL OUTCOMES OF RESPONSE TIME (MS)

(x-axis) – Total Load of IoTdevices(Mbps)	Response Time (ms)- (y-axis)		
	IFC	ARF	Proposed
100	85	75	65
300	95	90	85
500	130	110	100
700	120	110	105
900	140	130	120

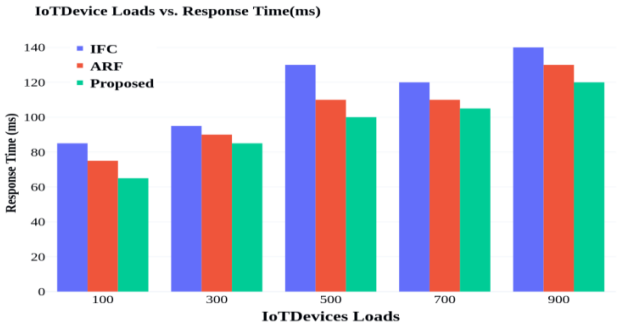


Fig. 4. IoTdevices vs. Response time (ms).

The numerical results display the response time (in ms) of three models, IFC, ARF, and the proposed, for different loads of IoT devices (in Mbps) in Fig. 4 and Table III. For a 100 Mbps load, response times are 85 ms for IFC, 75 ms for ARF, and 65 ms for the proposed method, with a 23.5% reduction from IFC. As the load is increased to 300 Mbps, the proposed method has improved efficiency at 85 ms, as opposed to 95 ms (IFC) and 90 ms (ARF). At 500 Mbps, the proposed method's response time is 100 ms, 23% quicker than IFC (130 ms). For 700 Mbps, the proposed method has 105 ms, representing a 12.5% improvement over IFC (120 ms). At the maximum load of 900 Mbps, the proposed technique reaches 120 ms with a substantial decrease in delay from IFC (140 ms) and ARF (130 ms), validating its better performance at all levels of load.

2) Total load (Mbps) vs. Throughput (Mbps): The correspondence between Total Load (Mbps) and Throughput (Mbps) within a smart IoT system can be expressed as:

$$Th = TL \times (1 - PL) \quad (44)$$

where,

Th is the Throughput (Mbps) (the actual data successfully delivered).

L is the Total Load (Mbps) (rate of incoming data).

P is the Packet Loss Ratio (ratio of lost packets).

As the Total Load *TL* rises, throughput *Th* rises in a linear manner until the system is fully utilized. After that, congestion creates packet loss *PL*, diminishing the effective throughput.

TABLE IV. NUMERICAL OUTCOMES OF THROUGHPUT (MBPS)

(x-axis) – IoTDevices Load (Mbps)	Throughput (Mbps)- (y-axis)		
	IB-SEEC	ARF	Proposed
100	95	97	99
300	270	285	297
500	445	470	492
700	615	648	685
900	790	835	878

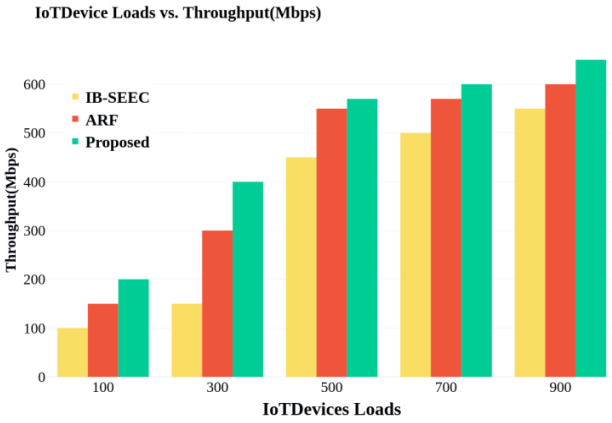


Fig. 5. IoTDevices loads vs. Throughput (Mbps).

The numerical results show the throughput (in Mbps) of three models, IB-SEEC, ARF, and the proposed method, for various IoT device loads (in Mbps) in Fig. 5 and Table IV. For a 100 Mbps load, the throughput of IB-SEEC is 95 Mbps, ARF is 97 Mbps, and the proposed method is 99 Mbps. At 300 Mbps load, the proposed method achieves 297 Mbps, compared with 270 Mbps for IB-SEEC and 285 Mbps for ARF. At 500 Mbps, the proposed method reaches 492 Mbps, outperforming IB-SEEC (445 Mbps) and ARF (470 Mbps). At 700 Mbps, the proposed method delivers 685 Mbps, compared with 615 Mbps for IB-SEEC and 648 Mbps for ARF. At the peak load of 900 Mbps, the proposed method achieves 878 Mbps, while IB-SEEC and ARF achieve 790 Mbps and 835 Mbps, respectively, confirming the superior throughput performance of the proposed method across all load levels.

3) Number of epochs vs. Attack detection (%): The correlation between Number of epochs and Attack detection (%) in an intelligent IoT framework based on AI is:

$$AD(\%) = AD_{max} \times (1 - e^{-\lambda NE}) \quad (45)$$

where,

AD(%) is Attack Detection Rate (percentage of attacked cases detected).

AD_{max} is the Maximum possible Detection Accuracy by the model.

NE is the number of epochs (number of iterations taken during training).

λ is Learning Rate Constant (accounts for how rapidly accuracy increases with epochs).

As the Number of epochs (*NE*) grows, the Attack Detection Rate closes in on its maximum limit AD_{max}, and increasing epochs produces smaller and smaller improvements.

TABLE V. NUMERICAL OUTCOMES OF ATTACK DETECTION (%)

(x-axis) –Number of epochs	Attack detection (%)- (y-axis)		
	GARL	ARF	Proposed
10	15	25	35
20	20	30	45
30	40	55	65
40	68	75	85
50	75	85	95

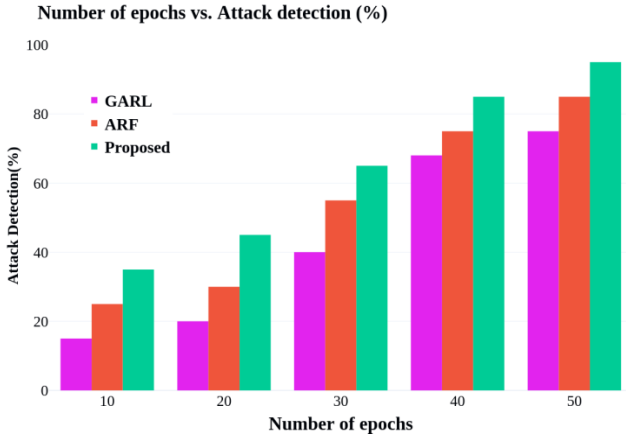


Fig. 6. Number of epochs vs. Attack detection (%).

The numerical results display the accuracy of attack detection (in %) of three models, namely, GARL, ARF, and the proposed method, on higher epochs in Fig. 6 and Table V. The proposed method performs 35% at 10 epochs, which is better than GARL (15%) and ARF (25%). As the epochs reach 20, the proposed method performs 45%, keeping a huge margin over GARL (20%) and ARF (30%). At 30 epochs, the proposed method is 25% more accurate than GARL (40%) and 10% more accurate than ARF (55%) at 65% accuracy. At 40 epochs, the proposed method still surpasses GARL (68%) and ARF (75%)

with 85% detection. At 50 epochs, the proposed method has the highest detection rate of 95%, outperforming GARL (75%) and ARF (85%), proving better accuracy and convergence speed at all epochs.

4) Number of epochs vs. false positive rate: The correlation between the Number of epochs and the False Positive Rate (FPR) in a smart IoT system with AI can be expressed as:

$$FPR = FPR_0 \times e^{-\varphi NE} \quad (46)$$

where,

FPR is the False Positive Rate (%).

FPR_0 is the Initial False Positive Rate at initial training.

NE is the Number of epochs (iterations during training).

φ is a Decay Constant (specifies how rapidly the false positive rate reduces).

As the number of epochs (NE) gets larger, the False Positive Rate lowers exponentially because the learning is becoming better, yet the decrease occurs at a slowing rate as it converges.

TABLE VI. NUMERICAL OUTCOMES OF FPR

(x-axis) –Number of epochs	FPR- (y-axis)		
	GARL	ARF	Proposed
10	0.08	0.07	0.05
20	0.07	0.06	0.04
30	0.06	0.05	0.035
40	0.05	0.04	0.03
50	0.04	0.03	0.02

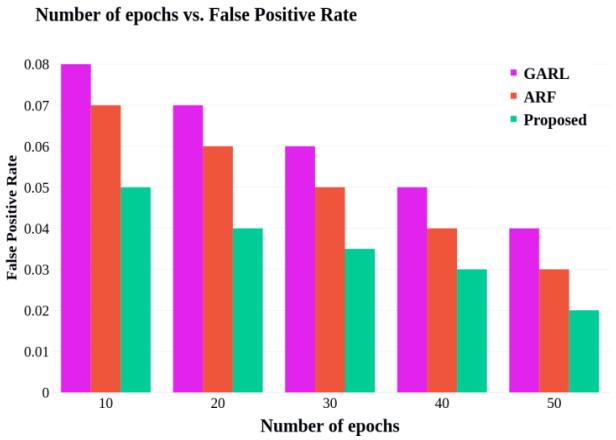


Fig. 7. Number of epochs vs. False positive rate.

The numerical results indicate the False Positive Rate (FPR) of three methods, GARL, ARF, and the proposed approach, against higher epochs in Fig. 7 and Table VI. With 10 epochs, the proposed approach has the lowest FPR of 0.05 compared to GARL (0.08) and ARF (0.07). With increased epochs to 20, the proposed approach also decreases the FPR to 0.04, whereas GARL and ARF have higher rates of 0.07 and 0.06, respectively. At 30 epochs, the proposed approach notes an FPR of 0.035, which is lower than GARL (0.06) and ARF (0.05). At 40 epochs, the proposed approach is still ahead with an FPR of 0.03, as against GARL (0.05) and ARF (0.04). For 50 epochs, the proposed model has the minimum FPR of 0.02, registering a considerable decline compared to GARL (0.04) and ARF (0.03), proving to be better in reducing false positives for all epochs.

5) Number of IoT devices vs. Energy consumption (%): The dependency between the Number of IoT devices and Energy Consumption (%) in a smart IoT system can be represented as:

$$EC = EC_{base} + NI \times PC_D \times AT \quad (47)$$

where,

EC is the Total Energy Consumption (%).

EC_{base} is the Base Energy Consumption (%) (System's idling energy consumption).

NI is the Number of IoT devices.

PC_D is the Power consumption per device (%).

AT is the Active time (hours) of each IoT device.

TABLE VII. NUMERICAL OUTCOMES OF ENERGY CONSUMPTION (%)

(x-axis)- Number of IoT devices	Energy Consumption (%)- (y-axis)		
	IB-SEEC	ARF	Proposed
10	0.9	0.8	0.5
20	1.0	0.7	0.6
30	1.3	0.9	0.8
40	1.4	1.0	0.9
50	1.5	1.3	1.0

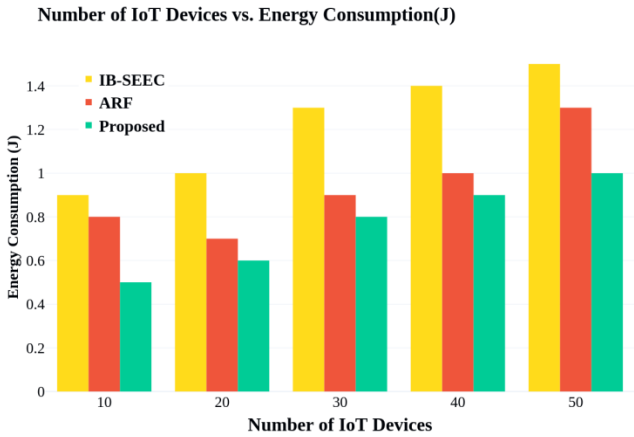


Fig. 8. Number of IoT devices vs. Energy consumption (%).

The numerical results show the energy usage (in %) of three methods, IB-SEEC, ARF, and the proposed method, over various numbers of IoT devices in Fig. 8 and Table VII. For 10 devices, the proposed method reflects the lowest energy usage at 0.5%, in contrast to IB-SEEC (0.9%) and ARF (0.8%). When the number of devices grows to 20, the proposed method reflects higher efficiency at 0.6%, as opposed to IB-SEEC and ARF with their 1.0% and 0.7%, respectively. At 30 devices, the proposed method reports 0.8%, which is lower than IB-SEEC (1.3%) and ARF (0.9%). At 40 devices, the proposed method reports 0.9%,

which is less than IB-SEEC (1.4%) and ARF (1.0%). At 50 devices, the proposed method is still the most effective at 1.0%, surpassing IB-SEEC (1.5%) and ARF (1.3%), with persistent energy savings at all levels of IoT devices.

To isolate the energy-saving contribution of Linear Network Coding (LNC), an ablation study was conducted by comparing three blockchain configurations under identical NS-3.35 settings: standard Hyperledger Fabric, Hyperledger Fabric with LNC, and the proposed integrated framework. At 50 IoT devices, energy consumption was 1.4% for standard Hyperledger Fabric, 1.1% for Hyperledger Fabric with LNC, and 1.0% for the proposed method. These results show that LNC reduces blockchain energy consumption by approximately 21.4% relative to standard Hyperledger Fabric, while the full proposed framework provides additional savings through coordinated load balancing and congestion control.

D. Research Summary

This study suggests an intelligent IoT architecture for a green city by incorporating Artificial Intelligence (AI) and information systems for maximizing resource management, improving public services, and urban efficiency. The research tackles major challenges like inefficient feature selection and extraction, load imbalance, Quality of Service (QoS) problems, and excessive blockchain energy consumption. The suggested smart IoT network includes 50 IoT devices, 2 base stations, 1 blockchain node, and 3 servers, with data gathered using the CICIDS2017 dataset. Data pre-processing is done using the Normalization technique to normalize data. Feature optimization is done through Wrapper-based Feature Selection and Autoencoder-based Feature Extraction, enhancing data relevance and model efficiency. QoS is ensured by a mechanism of dynamic load balancing via Software Defined Networking (SDN), and stateless Q-learning is applied to avoid congestion in data transmission. To minimize energy usage in blockchain functions, Linear Network Coding (LNC) is integrated into Hyperledger Fabric for minimizing energy usage in blockchain functions. A Quasi-Recurrent Neural Network (QRNN) with optimization from the Butterfly Optimization Algorithm (BOA) is utilized to detect cyber-attacks and improve detection rates, and reduce false positives. For execution, we utilize the ns3 Python simulation tool, and we also compare the graphs that were represented in Tables IV to VII, and Fig. 4 to 8. Our techniques' performance is investigated using numerical analysis, demonstrating that it performs better across all measures.

VI. Discussion

The NS-3.35 simulations modeled a smart-city IoT network with 50 devices, 2 base stations, 3 servers, and 1 blockchain node. Traffic was generated using a mixed UDP/TCP model with variable packet sizes and arrival rates ranging from 100 to 900 Mbps. Attack scenarios included probe, denial-of-service, and botnet traffic patterns derived from CICIDS2017 attack categories. Although the simulation scale is smaller than city-wide deployments, it provides a controlled environment for comparing load balancing, congestion control, and energy consumption trends.

Although stateless Q-learning does not maintain long-term historical state, it preserves short-term network feedback at each decision step through current utilization, queue length, and reward signals. This design reduces memory requirements while still enabling adaptive routing under changing traffic conditions.

The proposed framework indirectly supports urban sustainability by reducing response time, increasing throughput, lowering energy consumption, and improving cyber-attack detection accuracy. These improvements can support faster emergency response, more efficient resource usage, and more reliable public services in smart-city environments.

The integration of multiple intelligent modules introduces additional computational overhead compared with simpler single-module approaches. However, the proposed design mitigates this cost through feature selection, autoencoder-based dimensionality reduction, and efficient QRNN inference. The resulting performance gains in latency, throughput, security, and energy efficiency justify the added implementation complexity for city-scale IoT deployments.

Data privacy is preserved through Hyperledger Fabric authentication, encrypted transaction records, access control policies, and selective sharing of validated security events across system components. Sensitive raw traffic data remains localized within preprocessing and detection modules, while only authorized metadata and alerts are transmitted to the blockchain layer.

Failure resilience is supported through module-independent operation paths. If the blockchain layer becomes temporarily unavailable, attack detection and SDN-based traffic management continue using local logs and cached policies. If the attack detector fails, network traffic can still be routed using SDN and Q-learning rules, while blockchain logging resumes after service restoration.

Declaration of Generative AI Use: Generative AI tools were used only for grammar checking and minor language refinements during manuscript preparation. All research design, experiments, simulations, data analysis, and scientific conclusions were conducted manually by the authors.

VII. Conclusion

The suggested smart IoT framework effectively merges AI and information systems to enhance the sustainability and efficiency of smart city operations. Through the mitigation of key challenges like poor feature selection, inefficient load balancing, and excessive energy consumption, the framework guarantees efficient resource management, improved service delivery, and decreased system congestion. The integration of Wrapper-based Feature Selection and Autoencoder enhances data processing efficiency, while the SDN-based QoS mechanism and stateless Q-learning efficiently handle real-time load distribution and congestion prevention. In addition, the implementation of Hyperledger Fabric blockchain with Linear Network Coding (LNC) minimizes energy consumption, making blockchain operations sustainable. The combination of QRNN with BOA improves attack detection accuracy and reduces false positive rates, thus enhancing the security of IoT-based smart city systems. Performance analysis shows that the proposed system offers improved response times, increased

throughput, and lower energy consumption, thus making it a suitable solution for future sustainable smart city deployments. In summary, the work offers a robust, scalable, and effective framework that can be implemented to address the rising demands and challenges in smart city settings. Future research can extend the system to multi-city networks, enhance interoperability, and integrate more advanced AI models to achieve even higher efficiency and flexibility.

REFERENCES

[1] Auwalu, F. K., & Bello, M. (2023). Exploring the contemporary challenges of urbanization and the role of sustainable urban development: a study of Lagos City, Nigeria. *Journal of Contemporary Urban Affairs*, *7*(1), 175-188.

[2] Ali, I., & Rahman, A. (2024). Environmental Degradation: Causes, Effects and Solutions. International Journal for Multidisciplinary Research, 6(3), 1-10.

[3] Van Hoang, T. (2024). Impact of integrated artificial intelligence and internet of things technologies on smart city transformation. Journal of Technical Education Science, 19(Special Issue 01), 64-73.

[4] Stana, A., Toti, L., Kosova, R., & Prodani, F. (2019). Future of Durres: Smart City and Smart University. European Journal of Engineering and Technology Vol, 7(6).

[5] Barolli, A., Sakamoto, S., Bylykbashi, K., & Barolli, L. (2022). A hybrid intelligent simulation system for building IoT networks: Performance comparison of different router replacement methods for WMNs considering stadium distribution of IoT devices. Sensors, 22(20), 7727.

[6] Barolli, A., Sakamoto, S., Ozera, K., Barolli, L., Kulla, E., & Takizawa, M. (2018, February). Design and implementation of a hybrid intelligent system based on particle swarm optimization and distributed genetic algorithm. In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 79-93). Cham: Springer International Publishing.

[7] Rani, S., Mishra, R. K., Usman, M., Kataria, A., Kumar, P., Bhambri, P., & Mishra, A. K. (2021). Amalgamation of advanced technologies for sustainable development of smart city environment: A review. *IEEE Access*, *9*, 150060-150087.

[8] Hsu, C. H., Eshwarappa, N. M., Chang, W. T., Rong, C., Zhang, W. Z., & Huang, J. (2022). Green communication approach for the smart city using renewable energy systems. *Energy Reports*, *8*, 9528-9540.

[9] Dias, T., Fonseca, T., Vitorino, J., Martins, A., Malpique, S., & Praça, I. (2023, July). From data to action: Exploring AI and IoT-driven solutions for smarter cities. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 44-53). Cham: Springer Nature Switzerland.

[10] Stana, A., Golgota, A., & Toti, L. (2025). E-Mobility: First Step Towards the Future of a Smart City. Interdisciplinary Journal of Research and Development, 12(2 S1), 22-22.

[11] Ahmed, I., Zhang, Y., Jeon, G., Lin, W., Khosravi, M. R., & Qi, L. (2022). A blockchain-and artificial intelligence-enabled smart IoT framework for sustainable city. *International Journal of Intelligent Systems*, *37*(9), 6493-6507.

[12] Badidi, E. (2022). Edge AI and blockchain for smart sustainable cities: Promise and potential. Sustainability, 14(13), 7609.

[13] Das, D. K. (2025). Digital Technology and AI for Smart Sustainable Cities in the Global South: A Critical Review of Literature and Case Studies. *Urban Science*, *9*(3), 72.

[14] Das, D. K. (2024). Exploring the symbiotic relationship between digital transformation, infrastructure, service delivery, and governance for smart sustainable cities. Smart Cities, 7(2), 806-835.

[15] Simonofski, A., Vallé, T., Serral, E., & Wautelet, Y. (2021). Investigating context factors in citizen participation strategies: A comparative analysis of Swedish and Belgian smart cities. International Journal of Information Management, 56, 102011.

[16] Berigüete, F. E., Santos, J. S., & Rodriguez Cantalapiedra, I. (2024). Digital revolution: emerging technologies for enhancing citizen engagement in urban and environmental management. *Land*, *13*(11), 1921

[17] Ejaz, U., Ramon, W., & Olaoye, G. (2025). The Role of Big Data and AI in Smart Cities and Urban Planning.

[18] Neoaz, N. (2025). Internet of Things (IoT) and Smart Cities Examine how IoT technologies can improve urban living and infrastructure management. *Author Nahid Neoaz*.

[19] Dikshit, S., Atiq, A., Shahid, M., Dwivedi, V., & Thusu, A. (2023). The use of artificial intelligence to optimize the routing of vehicles and reduce traffic congestion in urban areas. *EAI Endorsed Transactions on Energy Web*, *10*, 1-13.

[20] Ismagilova, E., Hughes, L., Rana, N. P., & Dwivedi, Y. K. (2022). Security, privacy and risks within smart cities: Literature review and development of a smart city interaction framework. *Information Systems Frontiers*, 1-22.

[21] Mohammed, A. (2022). Cybersecurity in Smart Cities: Securing IoT and Smart Infrastructure. Journal of Innovative Technologies, 5(1).

[22] Alzyoud, F., Al-Falah, R., Tarawneh, M., & Tarawneh, O. (2024, March). Security challenges and solutions in smart cities. In International Conference on Advances in Computing Research (pp. 256-267). Cham: Springer Nature Switzerland.

[23] Vijarania, M., Gupta, S., Agrawal, A., Adigun, M. O., Ajagbe, S. A., & Awotunde, J. B. (2023). Energy efficient load-balancing mechanism in integrated IoT–fog–cloud environment. Electronics, 12(11), 2543.

[24] Yuvaraj, N., Praghash, K., Logeshwaran, J., Peter, G., & Stonier, A. A. (2023). An artificial intelligence based sustainable approaches—IoT systems for smart cities. In AI Models for Blockchain-Based Intelligent Networks in IoT Systems: Concepts, Methodologies, Tools, and Applications (pp. 105-120). Cham: Springer International Publishing.

[25] Nagarajan, S. M., Deverajan, G. G., Chatterjee, P., Alnumay, W., & Muthukumaran, V. (2022). Integration of IoT based routing process for food supply chain management in sustainable smart cities. Sustainable Cities and Society, 76, 103448.

[26] Jannu, S., Dara, S., Thuppari, C., Vidyarthi, A., & Gupta, D. (2023). An advanced energy management and harvesting system for network lifetime for industrial IoT in smart cities. IEEE Internet of Things Journal.

[27] Wen, W., Demirbaga, U., Singh, A., Jindal, A., Batth, R. S., Zhang, P., & Aujla, G. S. (2023). Health monitoring and diagnosis for geo-distributed edge ecosystem in smart city. IEEE Internet of Things Journal, 10(21), 18571-18578.

[28] Ajao, L. A., & Apeh, S. T. (2023). Secure edge computing vulnerabilities in smart cities sustainability using petri net and genetic algorithm-based reinforcement learning. Intelligent Systems with Applications, 18, 200216.

[29] Khadidos, A. O., Shitharth, S., Manoharan, H., Yafoz, A., Khadidos, A. O., & Alyoubi, K. H. (2022). An intelligent security framework based on collaborative mutual authentication model for smart city networks. IEEE Access, 10, 85289-85304.

[30] Luo, J., Zhao, C., Chen, Q., & Li, G. (2022). Using deep belief network to construct the agricultural information system based on Internet of Things. The Journal of Supercomputing, 78(1), 379-405.

[31] Kolhe, R. V., William, P., Yawalkar, P. M., Paithankar, D. N., & Pabale, A. R. (2023). Smart city implementation based on Internet of Things integrated with optimization technology. Measurement: Sensors, 27, 100789.

[32] Faisal, M., Sahabuddin, Hasiri, E. M., Darniati, Abd Rahman, T. K., Mulyadi, I., & Widia, I. D. M. (2025). Assessing AI-Driven Personalization in Smart Cities Using Hybrid Machine Learning and MCDM Approach. HighTech and Innovation Journal, 6(3), 770–792.

[33] Huy, P. Q., & Phuc, V. K. (2025). Impact of Technologically Vigilant Leadership on Smart Sustainable Circular Supply Chain Management. Emerging Science Journal, 9(4), 1868–1885.

[34] Xiao, Y., & Dong, Y. (2025). Optimizing AIGC Technology for IoT Devices with Deep Learning. HighTech and Innovation Journal, 6(3), 976–990.